Modelling Strategies for Functional Magnetic Resonance Imaging

Madsen, Kristoffer Hougaard

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Modelling Strategies for Functional Magnetic Resonance Imaging

Kristoffer H. Madsen

Kongens Lyngby 2008
IMM-PHD-2008-203

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN 0909-3192

Summary

This thesis collects research done on several models for the analysis of functional magnetic resonance neuroimaging (fMRI) data. Several extensions for unsupervised factor analysis type decompositions including explicit delay modelling as well as handling of spatial and temporal smoothness and generalisations to higher order arrays are considered. Additionally, an application of the natural conjugate prior for supervised learning in the general linear model to efficiently incorporate prior information for supervised analysis is presented. Further extensions include methods to model nuisance effects in fMRI data thereby suppressing noise for both supervised and unsupervised analysis techniques.

Resumé

Denne afhandling omhandler videnskabeligt arbejde i relation til modellering og analyse af data fra funktionel magnetisk resonans skanning (fMRI). I denne forbindelse er flere udvidelser af eksisterende modeller til eksplorativ dataanalyse blevet udviklet. Disse modeller inkluderer eksplicit modellering af tidsforsinkelser i faktoranalyse lignende modeller, regularisering af parametre til sikring af tidslig og rumlig jævnhed samt generalisering af disse modeller til højere ordens data (tensorer). Herudover præsenteres muligheden for at indbygge ekstra information i klassiske analyser ved hjælp af den naturlige konjugerede prior for den generelle lineære model. Desuden omhandler afhandlingen modellering af forstyrrende signaler, som kan tjene til at undertrykke støj i både klassiske og eksplorative analyser af fMRI data.

Preface

This thesis was prepared at Informatics Mathematical Modelling (IMM), the Technical University of Denmark (DTU) in partial fulfillment of the requirements for acquiring the Ph.D. degree in mathematical modelling.
A substantial part of the research was done at the Danish Research Centre for Magnetic Resonance (DRCMR) at Copenhagen University Hospital and either related to or part of projects running at DRCMR.
The project was funded by DTU and supervised by Professor Lars Kai Hansen with additional supervision provided by the three co-supervisors: Ph.D. Torben Ellegaard Lund (TEL), Ph.D. Karam Sidaros (KS) and Professor Axel Larsen (AL). TEL was employed at DRCMR until fall 2007 where he moved to the Center of Functionally Integrative Neuroscience at Aarhus University Hospital whereas KS has been employed at DRCMR during the entire Ph.D., however, he moved to an administrative position during the last two years. AL from the Center for Visual Cognition (CVC) at the University of Copenhagen was much involved in the research related to visual function done in the beginning of the Ph.D. before he was appointed head of the department of psychology at the University of Copenhagen.
In addition, I have had the pleasure of working with numerous people with different backgrounds and as such the diversity of these people have directed my research to many interesting topics of which only some are covered in this thesis. The topics covered within this thesis are related to machine learning applied to neuroimaging data in particular modelling of signals acquired using fMRI.

The thesis consists of a summary report and a collection of selected research papers written during the Ph.D. published or submitted for publication elsewhere.

Kgs. Lyngby, 2008

Kristoffer Hougaard Madsen

Acknowledgements

I thank my supervisor, Lars Kai Hansen, for guidance and inspiration. In addition I also owe many thanks to my co-supervisor, Torben E. Lund, for great inspiration, for always being willing to discuss various research topics and not least for his enjoyable company and friendship during the Ph.D.
I wish also to thank Olaf Paulson who on behalf of DRCMR provided the possibility of using the MR scanner when needed, and a warm welcome at DRCMR including academic and financial support.
I would also like to thank co-workers at both the ISP group at IMM and DRCMR at Hvidovre Hospital for their encouragement and company during the last three years. In particular I wish to thank Morten Mørup, without whom this Ph.D. would not have been possible in its current form and for opening my eyes to the exciting topic of multi-way analysis. I have really enjoyed our friendship and many fruitful discussions on almost every topic covered in this thesis.
I also wish to thank other people with whom I have had the pleasure to work with during the Ph.D., in particular this includes Daniel Jakobsen, Kirsten Nielsen, Mark Schram Christensen, Karl Sjöstrand, Finn Aarup Nielsen and Arnold Skimminge for their expertise and pleasant company.
In addition I wish to thank Anne-Marie Dogonowski, Astrid Rosenstrand Lou, Dan Milea, Daniel-Robert Chebat, Jon Wegener, Maurice Ptito and Xingchen Wu for working together with me on various interesting topics unfortunately not covered in this thesis.
Special thanks are directed to friends (especially Mark Porsby for proof-reading of the manuscript), co-workers and family for help and guidance throughout the project period, in particular I thank my family, Birgitte and Fiona, for their love and support. Last but not least special thanks go to the anonymous subjects for participation, concentration, patience, tolerance and not moving too much during the occasionally boring and exhausting data acquisition sessions performed during the Ph.D.

Publications

Journal Papers

[B.1] Kristoffer H. Madsen and Lars Kai Hansen. Efficient Bayesian detection of fMRI signals in temporally correlated noise. *NeuroImage*. Submitted.

[B.2] Morten Mørup, Lars Kai Hansen, Sidse M. Arnfred, Lek-Heng Lim, Kristoffer H. Madsen. Shift Invariant Multilinear Decomposition of Neuroimaging Data. *NeuroImage*. Accepted.

[B.3] Kristoffer H. Madsen, Lars Kai Hansen, Morten Mørup Combining Time and Frequency Domain Optimization: Shifts, Convolution and Smoothness in Factor Analysis Type Decompositions. *Journal of Signal Processing Systems special issue on MLSP2007*. Submitted.

[•] Kirsten Korsholm, Kristoffer H. Madsen, Jette L. Frederiksen, James B. Rowe, Torben E. Lund Cortical neuroplasticity in patients recovering from acute optic neuritis. *NeuroImage*. Accepted.

[•] Daniel J. Jakobsen, Lars Kai Hansen, Kristoffer H. Madsen. Bayesian Model Comparison in Nonlinear BOLD fMRI Hemodynamics. *Neural Computation*, volume 20, 2008, number 3, pages 738–755.

[B.6] Kirsten Korsholm, Kristoffer H. Madsen, Jette Lautrup Frederiksen, Arnold Skimminge, Torben E. Lund. Recovery from optic neuritis: an ROI-based analysis of LGN and visual cortical areas. *Brain*, volume 130, 2007, pages 1244–1253.

[•] Torben E. Lund, Kristoffer H. Madsen, Karam Sidaros, W. L. Luo, Tom E. Nichols. Non-white noise in fMRI: does modelling have an impact? *NeuroImage 29*, volume 18, number 7, 2006, pages 54–66.

[B.7] Axel Larsen, Kristoffer H. Madsen, Torben E. Lund, Claus Bundesen. Images of Illusory Motion in Primary Visual Cortex. *Journal of Cognitive Neuroscience*, volume 18, number 7, 2006, pages 1174–1180.

[•] Finn A. Nielsen, Mark S. Christensen, Kristoffer H. Madsen, Torben E. Lund, Lars Kai Hansen. fMRI neuroinformatics. *IEEE Engineering in Medicine and Biology Magazine*, volume 25, 2006, pages 112–119.

[•] Mark Schram Christensen, Thomas Z. Ramsøy, Torben E. Lund, Kristoffer H. Madsen, James B. Rowe. An fMRI study of graded visual perception. *NeuroImage 31*, 2006, pages 1711–1725.

[•] Kirsten Nielsen, Kristoffer H. Madsen, Jette Lautrup Frederiksen, Anne Mette Leffers, Torben E. Lund. Letter to the Editor: Functional magnetic resonance imaging corresponds to Humphrey perimetry in a patient with pituitary adenoma. *Acta Ophthalmologica Scandinavica*, volume 84, 2006, volume 25, pages 267–268.

Peer Reviewed Conference Papers / Abstracts (selected)

[•] Morten Mørup, Kristoffer H. Madsen, Lars Kai Hansen. Approximate L0 constrained Non-negative Matrix and Tensor Factorization. *ISCAS 2008*, special session on Non-negative Matrix and Tensor Factorization and Related Problems, 2866.

[B.4] Morten Mørup, Kristoffer H. Madsen, Lars Kai Hansen. Shifted Non-negative Matrix Factorization. *IEEE Machine Learning for Signal Processing 2007*.

[B.5] Morten Mørup, Kristoffer H. Madsen, Lars Kai Hansen. Shifted Independent Component Analysis. *7th International Conference on Independent Component Analysis and Signal Separation*, 2007 pages: 89-96, 2007

[•] Lars Kai Hansen, Kristoffer H. Madsen, Tue Lehn-Schiøler. Adaptive regularization of noisy linear inverse problems. *EUSIPCO 2006*.

[•] Daniel J. Jacobsen, Kristoffer H. Madsen, Lars Kai Hansen. Identification of non-linear models of neural activity in BOLD fMRI. *3rd IEEE International Symposium on biomedical Imaging: Macro to Nano*, 2006, pages 952-955.

[•] Karl Sjöstrand, Torben E. Lund, Kristoffer H. Madsen, Rasmus Larsen. Sparse PCA, a new method for unsupervised analyses of fMRI data. *Proceedings, International Society of Magnetic Resonance In Medicine 2006*.

[B.8] Kristoffer H. Madsen, Torben E. Lund. Unsupervised Modelling of Physiological Noise Artifacts in fMRI Data, *Proceedings, International Society of Magnetic Resonance In Medicine 2006*

[•] Kirsten Nielsen, Kristoffer H. Madsen, Jette Lautrup Frederiksen, Torben Ellegaard Lund. fMRI used as a diagnostic tool. *Human Brain Mapping 2005*.

[B.9] Kristoffer H. Madsen, Torben E. Lund. Simultaneous acquisition of polar and eccentricity mappings of the human visual cortex using fMRI, *Proceedings, International Society of Magnetic Resonance In Medicine 2005*

Contents

List of Abbreviations

ALS	Alternating Least Squares
BLUS	Best Linear Unbiased with a Scalar diagonal covariance matrix
BOLD	Blood Oxygenation Level Dependent
BPD	Basis Pursuit De-noising
BSS	Blind Source Separation
$CMRO_2$	Cerebral Metabolic Rate of O_2
CSF	Cerebral Spinal Fluid
DFT	Discrete Fourier Transform
dHb	deoxygenated Haemoglobin
EEG	ElectroEncephaloGraphy
EM	Expectation-Maximisation
EPI	Echo Planar Imaging
FA	Factor Analysis
FDR	False Discovery Rate
FFT	Fast (discrete) Fourier Transform – Algorithmic procedure to perform the Discrete Fourier Transform in $\mathcal{O}(n \log n)$ operations
GLM	General Linear Model
GLS	Generalised/weighted least squares
GRF	Gaussian Random Fields
Hb	Haemoglobin
HRF	Haemodynamic Response Function
ICA	Independent Component Analysis
i.i.d.	Independently identically distributed
IRLS	Iterative Re-weighted Least Squares
LASSO	Least Absolute Shrinkage and Selection Operator
LTI	Linear Time Invariant
MEG	MagnetoEncephaloGraphy

MNI305 .... Montreal Neurological Institute template based on MRI scans of 305 individuals
MRI ..... Magnetic Resonance Imaging
MR ..... Magnetic Resonance
MU ..... Multiplicative Update
NIG ..... Normal-Inverse Gamma
NIW ..... Normal-Inverse Wishart
NMF ..... Non-negative Matix Factorisation
NR ..... Nuisance Region
NVR ..... Nuisance Variable Regression
OLS ..... Ordinary Least Squares
PACE ..... Prospective Acquisition CorrEction
PCA ..... Principal Component Analysis
PET ..... Positron Emission Tomography
SC ..... Sparse Coding
SNR ..... Signal to Noise Ratio
SPMd ..... Statistical Parametric Mapping diagnostic
SVD ..... Singular value decomposition
$T_2^\star$ ..... Transversal relaxation time in presence of inhomogeniety
TFGM ..... Time-Frequency Gradient Method
uNVR ..... Unsupervised Nuisance Variable Regression

Notation

x	Scalar
$\imath$	Imaginary unit $\sqrt{-1}$
x^*	Complex conjugate of x
$\boldsymbol{x}$	Vector
x_i	i^{th} element of vector $\boldsymbol{x}$
$\boldsymbol{X}$	Matrix or two-way array
$\boldsymbol{x}_i$	i^{th} column of matrix $\boldsymbol{X}$
x_{ij}	Element from i^{th} row and j^{th} column of matrix $\boldsymbol{X}$
$\boldsymbol{I}$	Diagonal matrix (ones in main diagonal zeros elsewhere)
$\mathcal{X}$	Tensor/Hypermatrix of more than two dimensions
$\boldsymbol{X}_{(n)}$	The n-mode matricising operation: $\mathcal{X}^{I_1 \times I_2 \times \ldots \times I_N} \rightarrow \boldsymbol{X}^{I_n \times I_1 I_2 \ldots I_{n-1} I_{n+1} \ldots I_N}$
$\boldsymbol{X} \otimes \boldsymbol{Y}$	Kronecker product, i.e. $\boldsymbol{X} \otimes \boldsymbol{Y} = \begin{bmatrix} x_{1,1}\boldsymbol{Y} & \ldots & x_{1,j}\boldsymbol{Y} \\ \vdots & \ddots & \vdots \\ x_{I,j}\boldsymbol{Y} & \ldots & x_{I,j}\boldsymbol{Y} \end{bmatrix}$
$\boldsymbol{X} \odot \boldsymbol{Y}$	Khatri-Rao product, i.e. $\boldsymbol{X} \odot \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{x}_1 \otimes \boldsymbol{y}_1 & \boldsymbol{x}_2 \otimes \boldsymbol{y}_2 & \ldots & \boldsymbol{x}_J \otimes \boldsymbol{y}_J \end{bmatrix}$
$\times_n$	n-mode tensor product, $(\mathcal{X} \times_n \boldsymbol{Y})_{i_1,i_2,\ldots,j_n,\ldots,i_N} = \sum_{i_n} x_{i_1,i_2,\ldots,i_n,\ldots,i_N} y_{j_n,i_n}$
$\lvert x \rvert$	Absolute value of scalar x
$\lVert \boldsymbol{X} \rVert_F$	Frobenius norm of $\boldsymbol{X}$, i.e. $\lVert \boldsymbol{X} \rVert_F = \sqrt{\sum_{i,j} x_{i,j}^2}$
$\lVert \boldsymbol{X} \rVert_1$	l_1 norm of $\boldsymbol{X}$, i.e. $\lVert \boldsymbol{X} \rVert_F = \sqrt{\sum_{i,j} \lvert x_{i,j} \rvert}$
$\mathcal{F}(\boldsymbol{x}) = \widetilde{\boldsymbol{x}}$	Discrete Fourier transform of vector $\boldsymbol{x}$
$\mathcal{F}^{-1}(\widetilde{\boldsymbol{x}})$	Inverse discrete Fourier transform of vector $\widetilde{\boldsymbol{x}}$
$\boldsymbol{X}^\dagger$	Moore-Penrose generalised/pseudo inverse of matrix $\boldsymbol{X}$
$\mathrm{E}(x)$	Expected value of x
$\mathrm{var}(x)$	Variance of x
$\lvert \boldsymbol{X} \rvert$	Determinat of the matrix $\boldsymbol{X}$
$\mathrm{Tr}(\boldsymbol{X})$	Trace of the matrix $\boldsymbol{X}$
$\nabla_\theta X$	Gradient of X with respect to θ
$f'(x) = \frac{\partial f}{\partial x}$	Derivative of the function f with respect to the variable x

Chapter 1

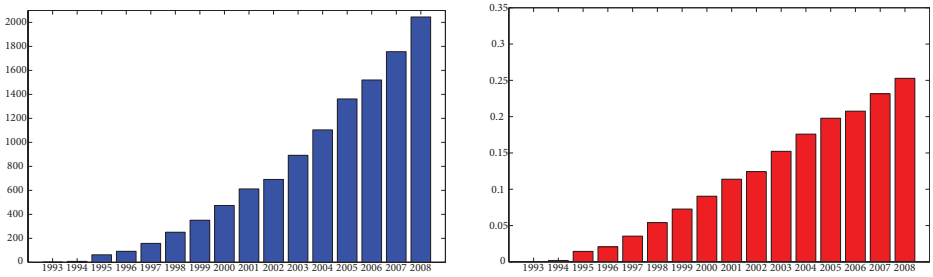
Introduction

1.1 Neuroimaging

Neuroimaging refers to directly or indirectly image the structure and/or function of the brain. In general imaging techniques can be divided into two broad categories: structural and functional imaging. Structural imaging refers to the observation/study of the anatomical features of the brain, whereas functional imaging refers to the study of brain function.

This thesis is concerned with applications of machine learning in the field of of functional imaging in particular the functional magnetic resonance imaging technique (fMRI).

Ever since its discovery fMRI (Ogawa *et al.*, 1990a) has received an increasing interest which can be demonstrated by the vast increase in publications on fMRI, figure 1.1. In particular fMRI has become an important tool in the efforts to advance our knowledge on brain function. Interpretation of data obtained from fMRI experiments is challenging due to limited knowledge about the origins of the signal, low signal to noise ratio and the massive amount of data generated. In this thesis noise reduction, supervised- and unsupervised analysis techniques are applied in attempts to obtain more parsimonious representations that facilitate understanding and overview of the data.



(a) fMRI publications (b) fraction of fMRI publications (%)

Figure 1.1: The figure shows how the amount of fMRI publications recorded in the online data base PUBMED (`http://www.pubmed.org`) has evolved since 1993. The numbers illustrated in the left panel are the number of PUBMED publications per year containing fMRI in either the title or the abstract. For the last year (2008) the number of publications are based on linear extrapolation from the 23rd of April. Note that the amount of publication in 2007 amounted to approximately 35 per week, indicating that the fMRI is a highly active research area. In the right panel the fraction of publications on fMRI is illustrated by normalising with the total number of publications recorded in PUBMED per year. It is seen that there is a substantial increase in the number of publications per year even when correcting for the increase in the total number of publications. Note that the numbers may be somewhat inaccurate, however, they large agree with numbers reported in a recent review article by Bandettini (2007) commenting on the current challenges in fMRI.

1.2 Overview

The subjects covered in this thesis serve mainly to collect some of the research done during the Ph.D. and to introduce the papers attached in appendix B.

Chapter 2 – Background serves as a brief and simple introduction to interpreting brain images and functional neuroimaging. A full description of fMRI is well beyond the scope of this text, therefore readers who are not familiar with these subjects should refer to the literature for more information. A good introduction to magnetic resonance and fMRI are available in Buxton (2002) whereas a more detailed description of the echo-planar imaging (EPI) technique can be found in Schmitt *et al.* (1998).
Chapter 3 – Pre-processing describes typical preprocessing steps for fMRI and anatomical MRI data.
Chapter 4 – Data Analysis and Models introduces the general linear model as well as the factor analysis model and extension to multi-way arrays in the form of canonical decomposition/parallel factor analysis.
Chapter 5 – New Models introduces models developed during the Ph.D. including shift invariant multi-way analysis, smooth and sparse coding, Bayesian general linear model with a natural conjugate prior and slice-wise modelling of fMRI data. Readers familiar with the subjects of neuroimaging and analysis fMRI data are encouraged to skip directly to this section.

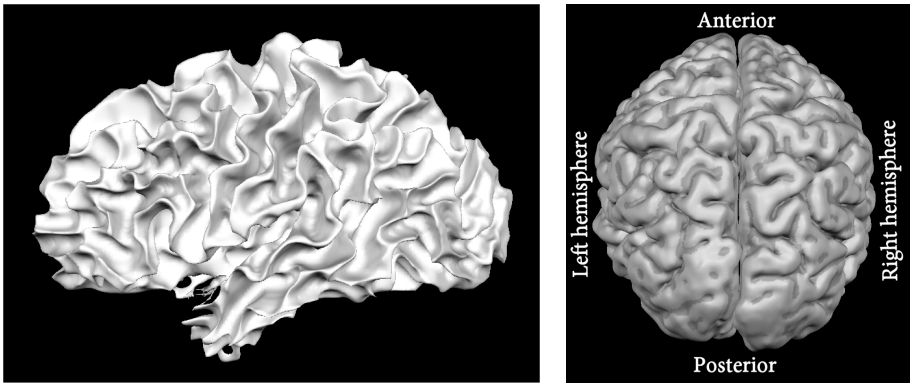
Chapter 2

Background

> Aristotle was famous for knowing everything. He taught that the brain exists merely to cool the blood and is not involved in the process of thinking. This is true only of certain persons.
>
> *William Jacob "Will" Cuppy*
> *(August 23, 1884 – September 19, 1949)*

2.1 Navigating the Brain

In a very coarse and superficial description the brain consists of white matter (WM), mostly constituted by myelinated axons, surrounded by grey matter (GM) containing cell bodies. This overall structure is illustrated in figure 2.1. In practice this very coarse description is far from correct as many (very important) GM structures are hidden inside the WM. Furthermore, the hard division into WM and GM is hardly appropriate. Figure 2.2 shows how the cerebral cortex can be divided into lobes. For examples the role of the occipital lobe is mainly related to processing of visual input whereas especially the frontal lobe is important for higher order cognitive functions. The structure of the brain is inherently three dimensional, however, it is possible to visualise the brain in isometric plane projections. For these types of projections the typically used convention for MRI



(a) White Matter (b) Grey matter

Figure 2.1: WM/GM **and hemispheres.** Panel (a) shows a reconstruction of the WM surface and panel (b) shows the GM surrounding it. Anterior represents frontal areas whereas posterior designates parts to the back. Inferior refers to bottom while superior refers to the top. Again medial are used to describe areas near the thalamus in the midbrain (division between hemispheres). Dorsal/ventral refers to over/below.

images is radiological convention (meaning the left hemisphere of the brain is displayed to the right, equivalent to viewing axial slices from the bottom). In the neuroimaging literature, however, neurological convention (left hemisphere displayed to the left) is more common. A projection obtained by a vertical cut parallel to the ears is denoted a sagittal plane, whereas a plane parallel to the face is a coronal plane. Finally, the projection obtained by a horizontal cut is denoted a transversal plane, three dimensional views of these cuts are illustrated in figure 2.3.

2.2 Functional Magnetic Resonance Imaging

The main focus of this thesis is on modelling of the blood oxygenation level dependent (BOLD) signal often used to achieve functional contrast in MRI. However, the usefulness of many of the methods addressed in this thesis should extend to other functional contrasts as well as other modalities such as positron emission tomography (PET), magnetoencephalography (MEG), electroencephalography (EEG) and optical imaging.

The description of fMRI will here be restricted to the BOLD signal even though

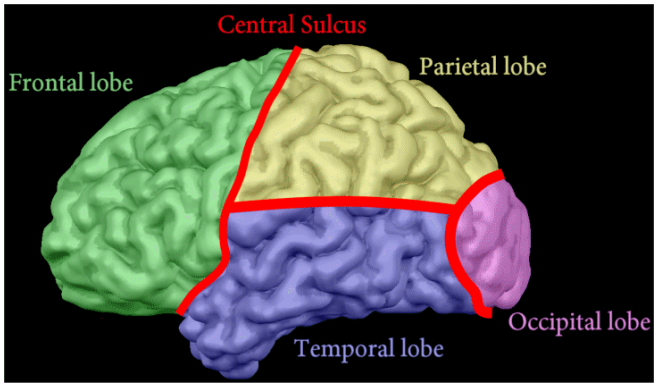
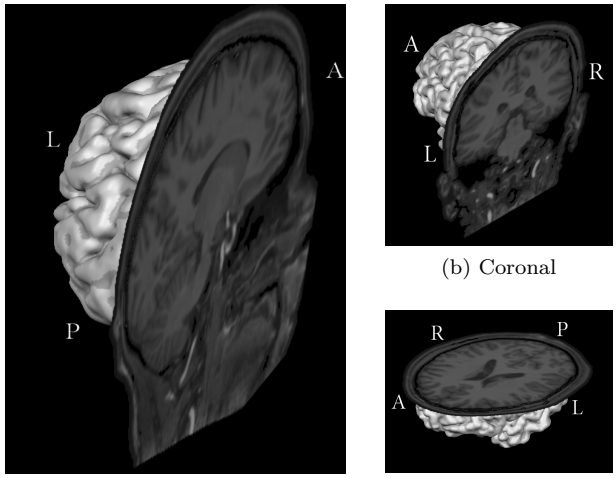


Figure 2.2: **Lobes in the cerebral cortex.** The figure is a reconstruction of the left GM surface. The different lobes have been roughly sketched and coloured. Gyri (singular gyrus) are the convex folds in the brain tissue, whereas areas with negative curvature (concave folds - dark) are in general called sulci (singular sulcus). Very deep sulci, such as the division between the hemispheres, are denoted fissures.

many different functional contrast measures exist. Other methods include arterial spin labelling (ASL), this technique relies on measuring blood perfusion by assessing the difference between a measurement where the blood (water) is tagged magnetically and a control experiment without tagging. The perfusion signal as measured by ASL is expected to be more directly related to neural activity than the BOLD signal and more stable over time (some sequences can even measure absolute perfusion) being a differential measure. The disadvantages, however, include lower signal to noise ratio, longer repetition times, and difficulties in getting full brain coverage. Other examples include for example the vascular-space-occupancy VASO technique for measuring blood volume changes (Lu *et al.*, 2003) and functional measurements by diffusion MRI, where water is expected to diffuse into the cell (cell swelling) in response to neuronal activation (Song *et al.*, 1996; Darquie *et al.*, 2001; Le Bihan, 2003; Li and Song, 2003; Song *et al.*, 2003; Le Bihan *et al.*, 2006).

2.2.1 Blood Oxygenation Level Dependent Signal

Although there exist other contrast measures which are expected to be more directly coupled to neuronal activation the BOLD signal remains the most widely used contrast for functional measurements in MR imaging. Advantages of this method include the relatively low complexity (in terms of measuring) and the



(a) Sagittal (b) Coronal (c) Transversal

Figure 2.3: **Isometric plane projections in the brain.** Panel (a) sagittal plane, panel (b) coronal plane, panel (c) transversal plane

high signal to noise ratio (SNR) when compared to competing methods. The BOLD is expected to be related to the local field potential as described (Logothetis *et al.*, 2001; Logothetis, 2002)) but the exact relation to neuronal activity is unknown. The main assumption behind this functional measure is (as is the case for many other methods) that increased neuronal activity gives rise to localised increased blood flow. BOLD measurements were first performed on rats by Ogawa *et al.* (1990b,a), and later on humans Ogawa *et al.* (1992).

Measurements of the BOLD signal is typically based on echo-planar imaging (EPI) techniques and the actual measured effect stems from signal loss due to presence of deoxygenated haemoglobin (dHb) (giving rise to field inhomogeneity). The sequences used are often based on gradient echo (GE) sequences due to the high SNR even though spin echo (SE) is expected to give better localisation (Bandettini *et al.*, 1994; Duong *et al.*, 2002). Oxygenated haemoglobin (Hb) is weakly diamagnetic (and has very little influence on the local field) whereas dHb is weakly paramagnetic causing shortening of the $T_2^\star$ relaxation time. The presence of dHb will therefore result in loss of MR signal (faster loss of signal).

Given these observations increased neuronal activity would be expected to cause increased oxygen consumption thereby decreasing the MR signal. However, the phenomenon is more complicated and in general the BOLD signal is expected to be caused by at least three effects; the cerebral blood flow (CBF), the cerebral

blood volume (CBV) and the cerebral metabolic rate of O_2 ($CMRO_2$).

The following description of the BOLD signal is motivated by the model suggested by (Buxton *et al.*, 1998), a mathematical description of the model will be given later in this section. An increased oxygen extraction due to neuronal activity results in a short increase in the dHb. This should give rise to a short decrease in the MR signal (initial dip), this effect however is rather small and may be difficult to observe in practice. The increased oxygen consumption gives rise to an increased CBF dominating the first effect thus decreasing dHb (increase in MR signal). The large increase in CBF is expected to be necessary because neuronal activity is limited by the blood flow in resting state. The increase in blood flow expand the blood vessels and increases CBV, thus increasing the amount of dHb. This effect is insignificant compared to the effect of CBF. However, it has been suggested that the post stimulus undershoot may be caused by CBV returning slower to baseline than CBF. In this context it should be noted that some contradicting experimental evidence exists (Frahm *et al.*, 2008). The total temporal extent of the BOLD response is expected to be approximately 30 seconds. In the case where the BOLD signal can be assumed to be linear time invariant (LTI) the output can be modelled as a simple convolution of an input signal (activation paradigm) driving the response (Boynton *et al.*, 1996). We will refer to the impulse response (used in convolution) as the canonical heamodynamic response function (HRF). An example of a HRF is displayed in figure 2.4.

Although there is much support that the BOLD signal is in general a non-linear function of the neuronal input, (Liu and Gao, 2000; Miller *et al.*, 2001; Kershaw *et al.*, 2001; Birn *et al.*, 2001; Huettel, 2004; Birn and Bandettini, 2005; Gu *et al.*, 2005; Wan *et al.*, 2006; Robinson *et al.*, 2006; Jin *et al.*, 2006), analyses of BOLD imaging data is typically based on linear models due to the attractive computational properties. A widely used example of such a model is the general linear model (GLM) introduced in section 4.4. Note that by including time derivatives (or similar expansions) in linear models it is possible to model variations in the response (delay and dispersion) and even (to some extent) account for non-linearities using Volterra expansions (Friston *et al.*, 1998).

In the following, however, we consider a more general case where the BOLD signal is described by a non-linear dynamical model formalising the observations given above in a mathematical framework. We assume that the BOLD signal $\boldsymbol{y}$ is governed by some function $g(\boldsymbol{x}, \boldsymbol{\theta})$ given the internal or external states (time varying) defined in the vector $\boldsymbol{x}$ and the in general unknown parameter vector $\boldsymbol{\theta}$

$$\boldsymbol{y}(t) = g(\boldsymbol{x}(t), \boldsymbol{\theta}),$$

where t denotes the (continuous) time variable. Several forms of the function

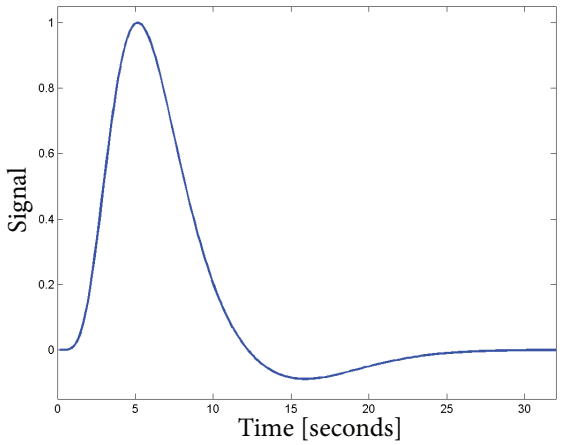


Figure 2.4: **Haemodynamic response function.** The figure illustrates the HRF used in the SPM5 software package based on a mixture of gamma functions. This HRF function is commonly denoted the SPM canonical HRF and describes the expected BOLD impulse response.

given above have been suggested (Friston *et al.*, 2000a; Buxton *et al.*, 2004), however, here we consider only the non-linear model suggested by (Buxton *et al.*, 1998), typically referred to as the original "Balloon model"

We start by simply stating that the change in blood volume, v over time is given by the difference between inflow $f(t)$ and the outflow $v^{1/\alpha}$ this last term is motivated by the viscoelastic windkessel model (α is an inverse stiffness parameter between zero and one)(Frank, 1930; Mandeville *et al.*, 1999), hence the term Balloon model,

$$\frac{\partial v(t)}{\partial t} = \frac{1}{\tau_0}(f(t) - v^{1/\alpha}) \tag{2.1}$$

where τ_0 is a time scaling parameter. We now consider the change in dHb which we assume to be determined by the inflow, outflow and the amount of oxygen extracted

$$\frac{\partial q(t)}{\partial t} = \frac{1}{\tau_0}(f(t)\frac{E_t}{E_0} - v^{(1/\alpha-1)}q(t)). \tag{2.2}$$

Here $E(t)$ is a function describing the extraction of oxygen often taken to be given by $E(t) = 1-(1-E0)^{1/f(t)}$. Where E_0 is the rest oxygen extraction fraction. The variable $s(t)$ describes a signal that depends on the neural activation (even though what exactly is meant by this notion is somewhat vague). We will

assume that this variable is governed by

$$\frac{\partial s(t)}{\partial t} = u(t) - s(t)/\tau_s - (f(t) - 1)/\tau_f, \tag{2.3}$$

where $u(t)$ is the input 'neural activity', τ_s and τ_f are time constants related to the dampening of $s(t)$ in absence of activity and inflow feedback respectively. We further assume that $s(t)$ control the inflow directly (linearly)

$$\frac{\partial f(t)}{\partial t} = s(t). \tag{2.4}$$

The remaining parameters (not described in terms of their time derivative) are assumed to be time stationary. Typically some of the parameters introduced are assumed to be known a priori (theoretically, or estimated from data in previous experiments). Other parameters can be estimated from the data; the main problem with this approach, however, is that the correct parameters are difficult to determine and reasonable priors for the parameters will have to be provided.

In Jacobsen *et al.* (2008) we compared a stationary version of the model above to an extended version of the Balloon model developed in Friston *et al.* (2000a) and Buxton *et al.* (2004) using a combination of Markov chain Monte Carlo and resampling techniques. Another comparison of these models using classical testing theory is presented in Deneux and Faugeras (2006) (based on ML estimates of the parameters and model likelihood).

2.2.2 Noise

Analysis of data obtained from BOLD fMRI are typically analysed in image-space (after reconstruction) considering only the magnitude of the complex valued signal even though some experimental evidence exist that the phase may provide additional information for localisation of effects (Rowe, 2005). Then noise in these magnitude images are expected to have Richian distribution (also known as the Rice distribution) assuming that the measurement noise is Gaussian (Gudbjartsson and Patz, 1995; Andersen, 1996; Sijbers *et al.*, 1998; den Dekker and Sijbers, 2005). Most analysis techniques make the often reasonable assumption that the distribution is Gaussian around a mean value.

Others sources of nuisance effects than the measurement error exist and in general we consider any unwanted contrast mechanism (i.e. not related to the effect of interest) as noise. A very prominent example is movement artefacts, the most obvious effect of movement is a simple rigid body displacement. In

practice, however, the effects of movement are much more complicated due to various other effects. First of all the MRI signal is not measured instantaneously but rather sampled by sequentially transversing the spatial frequencies complicating the artifacts in image space. Furthermore, multi-slice EPI acquisitions are typically acquired interleaved further complicating the situation. We briefly consider movement correction methods in section 3.3.

Low frequency fluctuation in BOLD imaging is often attributed to physiological effects such as spontaneous neural activation, however, the fact that it is also observed in phantoms and cadavers (Lund and Larsson, 1999; Smith *et al.*, 1999) indicate that effects are also caused by hardware instability (such as heating up of scanner hardware). This effect makes it particularly difficult to measure the BOLD signal over prolonged periods of time. A small part of this effect is caused by the steady slow decrease in the static magnetic field (B_0). However, in pratice this effect is drowned by other effects such as instability of the shim and radio frequency system, and in general drift may cause either increase or decrease of the signal. High-pass filtering prior to or during modelling is commonly used to reduce this problem when analysing BOLD fMRI data.

Other effects include physiological noise contributions due to cardiac pulsation, respiration and interaction with movement effects. We will describe these effects in further detail along with means of reducing the impact on data analysis in sections 4.5.1 and 5.5.

Chapter 3

Pre-processing

If you torture data sufficiently, it will confess to almost anything.

Fred Menger

3.1 Normalisation and Coregistration

Aligning and transforming brain volumes has been the topic of extensive research, and a full description is beyond the scope of this text, hence, only a selection of topics and references will be presented. We here make the distinction between coregistration and normalisation in that coregistration will refer to six-parameter rigid body transformation, whereas, normalisation will refer to more general image registration including non-linear deformation models. Coregistration can be formulated in terms of the more general linear/affine transformation

$$\boldsymbol{x}_{new} = \boldsymbol{A}\boldsymbol{x}, \tag{3.1}$$

where $\boldsymbol{x}$ is an vector describing the image position to be transformed augmented with 1 at the end, e.g. $[x_1\ x_2\ x_3\ 1]^T$ for the three dimensional case which we will consider in the following. Likewise, the vector $\boldsymbol{x}_{new}$ describes the transformed position. The affine (4×4) transformation matrix $\boldsymbol{A}$, in which the last row is constrained [0 0 0 1], describes a twelve parameter linear transformation,

e.g. translation, rotation, shearing and zooming. The affine transformation can be re-parameterised into a six-parameter transform describing only translation and rotation (see (Ashburner and Friston, 1997)) which forms the basis for coregistration or movement correction which we will consider in section 3.3. The six parameters are then optimised according to a cost function such as the sum of squared differences (only appropriate witin modality) or normalised mutual information with a template image. Unfortunately, the presence of local minima causes difficulties and global optimisation techniques such as multi-resolution approaches are often applied (Jenkinson *et al.*, 2002).

Automated normalisation methods can be roughly divided into two groups, either based on basis functions (known as low dimensional non-linear warping) or high dimensional non-linear warping. The basis function approach relies on defining a set of typical deformations for example based on a discrete cosine transform with frequencies up a selected cut-off frequency (Ashburner and Friston, 1999). The other class of methods typically rely on fluid deformation models or similar types of models (Christensen, 1994; Ashburner *et al.*, 1999).

One use of normalisation is the registration of an anatomical image to a standard template. The normalisation provides the possibility of identifying coordinates with respect to an anatomical template/atlas or direct investigation of the deformations (morphometry). A popular template is the widely used Talairach template (Talairach and Tournoux, 1988) based on detailed anatomical description from a post-mortem study of one brain. A more modern template is the Montreal Neurological Institute (MNI305) template based on the MRI scans of 305 young healthy right handed subjects (239 male, 66 female, mean age 23.4 ± 4.1) approximately in register with the Talaraich template (Evans *et al.*, 1993).

For more details on coregistration and normalisation procedures see Woods *et al.* (1992); Friston *et al.* (1995a); Ashburner and Friston (1997, 1999); Ashburner *et al.* (1999); Woods *et al.* (1998a,b); Jenkinson and Smith (2001); Jenkinson *et al.* (2002) and references herein.

3.2 Anatomical Data

Functional volumes can be acquired very fast (typically within 2-3 seconds), however, in terms of identifying anatomical features they have several shortcomings. In particular a high noise level, low spatial resolution and bad WM/GM contrast is usually to be expected. In order to accurately delineate anatomical features a high resolution T_1 weighted anatomical volume is usually acquired. By overlaying the activation from statistical analysis of functional scans, anatomi-

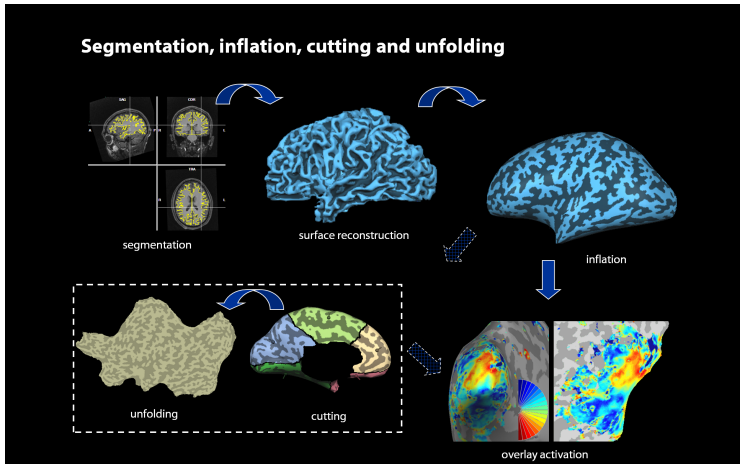


Figure 3.1: **Cortical surface mapping.** The figure show an overview of the steps involved in cortical surface mapping.

cal localisation of the effects is possible assuming that good alignment between the functional and anatomical scan can be provided using an appropriate co-registration method. The anatomical image also provides good GM/WM contrast making segmentation possible. Brain function is (to some extent) restricted to GM, largely located in a layer near the edges of the brain. This cortical GM represents a highly folded structure as seen in figure 2.1. Because activity is expected to be spatially extended along this folded structure it is sometimes preferable to display the activation on an unfolded representation of this cortical layer. We used this approach to visualise a perceptual filling in effect in Larsen *et al.* (2006). In the following we briefly summarise the steps involved in obtaining a surface based representation also known as cortical flattening. An overview of the steps are displayed in figure 3.1. It should be noted that normalisation is typically performed prior to the procedure described here.

- Inhomogeniety correction: Due to spatial variations in the sensitivity of the receiving head coil, tissue has different intensities according to the position within the head coil. This effect is especially pronounced for surface coil arrays. This causes the intensity of the same substance to vary when sampled at different spatial locations. This is particularly problematic for the segmentation into GM and WM as algorithms rely on differences in intensity to classify the tissue type. In order to correct for this difference in intensity a procedure know as coil inhomogeneity correction is typically performed prior to segmentation. These correction methods are typically based on the (reasonable) assumption that the variations in sensitivity

varies slowly over space. One often used approach is to estimate a three dimensional bias field based on low frequency spatial transforms (DCT, DFT) or slowly varying three dimensional Legendre polynomials to the image intensities within a rough initial segmentation (pre-segmentation) of the data (Vaughan *et al.*, 2001). More optimal procedures include the bias correction, segmentation and normalisation in the same estimation process (Ashburner and Friston, 1997, 2005; Zhang *et al.*, 2001).

- Segmentation: Segmentation refers to classifying voxels into WM,GM,cerebral spinal fluid (CSF) and possibly other tissue types. A full description of methods for classification is well beyond the scope of this text. However, in appendix A.1.6 a simple segmentation procedure providing a representation of the usually well-defined WM/GM boundary surface which will suffice of construction of a cortical surface representation.

  The tesselated volume based on the segmented volume can be deformed iteratively according to an local energy function. The energy function is constructed with a spring term smoothing the surface and an intensity based term attempting to move the surface to obtain a target intensity. Depending on whether the target intensity is based on WM/GM boundary intensity or the GM/CSF matter boundary intensity either the WM/GM surface or the GM/CSF surface (also known as the pial surface) can be obtained (Dale *et al.*, 1999).

- Inflation: The inflation process is performed in order to obtain a surface free from folds allowing easy visualisation or deformation in order to align the surface to a template. This process is usually performed by iteratively minimising an energy function penalising folds while attempting to keep local distances constant. Depending on the exact form of the energy function either an inflated surface as shown in figure 3.1 or a sphere can be obtained. The sphere representation is useful for aligning the surface to a surface based atlas (Fischl *et al.*, 1999a,b).

- Unfolding: In order to obtain an unfolded representation allowing easy visualisation surface cuts has to be made. For studies imaging the visual cortex a cut that approximately separate the occipital lobe from the rest of the brain and a cut along the calcarine sulcus is usually made. Finally, the surface is projected onto a plane (two sides), unfolding is now the process of warping the regions from the side on which the cuts were made onto the other side. The unfolding procedure typically introduce severe distortions in local distances on the surface as compared to the original mesh (especially in regions near the cuts). The distortion may be reduced by once again minimising a local length preserving energy function (now keeping

the desired two dimensional representation fixed) (Fischl *et al.*, 1999a). The resulting surface is a flat representation of the cortical GM where activation can be overlaid using the original positions from the folded mesh for each vertex.

3.3 Motion correction

Even minor movements during an fMRI experiment severely degrade the quality of the signal. The main reason is that some voxels will be moved to another part of the brain potentially with a different signal intensity (for example moving a WM voxel into GM or vice versa). For the same reason this effect is normally most prominent at edges in the images for example at intersection between CSF and GM/WM (for example near the ventricles and at the edge of the brain/head). The signal contamination due to movement may be reduced by the use of a six parameter rigid body transformation (displacement and rotation) (Friston *et al.*, 1995b; Woods *et al.*, 1992). By the use of this simple correction method we implicitly assume that movement happens between volumes. This assumption will most certainly not be satisfied because most sequences use almost all the time for either magnetisation or readout. In reality, the effects of movement are highly non-linear and include effects of both field inhomogeniety and spin history. Nevertheless, rigid body transformation greatly reduces the effects of movement and are virtually always applied.

Interpolation schemes such as either tri-linear, B-spline or sinc interpolation are applied in order to estimate the effects of sub-voxel movements. The six movement parameters are typically determined by ML estimation based on either minimising the sum of squares difference from a reference image. In certain cases movement correction procedures have been reported to cause spurious effects especially when considering the least squares objective even when no movement is present (Freire and Mangin, 2001). It may therefore be advantageous to consider alternative more robust cost functions as described in Freire *et al.* (2002). An example of an alternative cost function is the normalised mutual information.

The reference image is typically chosen to be the first or middle image in the time series. Often a two step procedure is formed where the images are aligned to the first image during the first pass and then aligned to a mean image (calculated using the transformation obtained in the first pass) in the second pass.

Prospective movement correction procedures implemented directly on the scanner hardware rely on estimating movement during scanning and correcting the

gradients according to the detected movement. An example of this method is the prospective acquisition correction (PACE) method described in Thesen *et al.* (2000). The main disadvantage of this procedure is that movement is not corrected until it is detected in the images which causes regular rigid body re-alignment to still be necessary. An improvement to the PACE method was described in Speck *et al.* (2006) where detection of movement was based on optical tracking.

In addition to rigid body re-alignment correction of geometric distortions caused by movement by field inhomogeniety interaction effects was suggested in Andersson *et al.* (2001).

After the rigid body re-alignment procedure residual motion effects may still severely affect the images and further analysis (Lund *et al.*, 2005). Further correction for these effects have been suggested in Friston *et al.* (1996). This method relies on modelling of the nuisance effects caused by movement using a Taylor series expansion of the fitted movement parameters and spin history effects (movement parameters from the previous volume). This method is described in details in section 4.5.1.

3.4 Slice Time Correction

Sequences used to obtain the desired $T_2^\star$ weighting for functional imaging are typically recorded one slice at a time resulting in different acquisition time for each slice. A complication of this time delay is that the slices are usually recorded interleaved to reduce cross-talk between slices. Ignoring the differences in timing may cause problems especially in the case of fast shifting (event related) paradigms. The methods for correction of slice acquisition time typically rely on adapting either the data or the model to account for differences in slice acquisition time. The methods that attempt to correct the data rely on temporal interpolation changing the data to make it appear as if all slices were acquired simultaneously. Interpolation in the time series requires that the signal is sufficiently sampled (meaning that the signal is band-limited and consist of frequencies lower than the Nyquist criterion). In fMRI this condition is often expected to be fulfilled for the signal of interest, however, physiological noise in particular cardiac induced noise is seldom sufficiently sampled compromising time interpolation. The other class of methods rely on changing the model to account for differences in acquisition time. In general, these fall into two categories. One category of methods attempt to make the model flexible enough to account for the differences in slice timing also allowing the model to capture differences in the BOLD response caused by differences in local physiology.

The other category adapts the model such that timing differences are modelled explicitly.

3.5 Smoothing

The strong traditions for spatial smoothing of data as part of the pre-processing exist for several reasons. Most importantly smoothing suppresses non spatially distributed noise by averaging neighbouring voxels. Also the smoothing process may help to reduce anatomical differences between subjects and improve normalisation by eliminating unnecessary fine details. Furthermore, the fact that Gaussian random field theory is often used to control the family-wise error in mass-univariate hypothesis tests for significant effects require that data exhibit smoothness over several millimeters to become in agreement with the assumptions usually made in Gaussian random field GRF theory Worsley *et al.* (1996). Drawbacks of smoothing include compromised spatial resolution. Smoothing is usually done by convolving with a three dimensional Gaussian kernel either in the frequency domain or approximately in the spatial domain to reduce computational cost.

In addition to spatial smoothing temporal smoothing/low pass filtering is also considered in some cases to suppress noise. This will introduce severe (known) temporal autocorrelation in the signal. If the introduced autocorrelation is expected to dominate over the autocorrelation in the data a framework where the known introduced autocorrelation is corrected can be formed as described in Worsley and Friston (1995). This can serve as a robust alternative to prewhitening, the two approaches are described in detail in (Friston *et al.*, 2000b).

3.6 High-Pass filtering

Due to the presence of low frequency fluctuations not related to the effects of interest (Lund and Larsson, 1999; Smith *et al.*, 1999) it is often desirable to remove low frequencies in the data prior to the analysis, this is typically performed by applying a high-pass filter. One way to perform this task is to include a low frequency Fourier or DCT expansion (Worsley and Friston, 1995) up to a desired cut-off frequency in the GLM used to model effects of interest. Other approaches use expansions in orthogonal polynomials (Worsley *et al.*, 2002) (this has the advantage that the resulting fit is free at both the end and the onset). The optimal cut-off frequency for the high pass filter should depend

on the specific scanner hardware, i.e. the optimal parameters could be found by analysing null time series data as done in (Lund *et al.*, 2006). During design of the stimulus paradigm it is important to consider the limitations caused by low frequency drift and ensure that the effects of interest occur at frequencies higher than the desired cut-off frequency.

Chapter 4

Data Analysis and Models

Statistical Analysis:

Mysterious, sometimes bizarre, manipulations performed upon the collected data of an experiment in order to obscure the fact that the results have no generalisable meaning for humanity. Commonly, computers are used, lending an additional aura of unreality to the proceedings.

Woodman (1979)

4.1 Temporal Autoregressive Modelling

We will denote the autoregressive (AR) model of order p an AR(p) model. In this model the signal y_t at time t is described in terms of the past signal values as given in the expression

$$y_t = \sum_{i=1}^{p} \lambda_i y_{t-i} + \varepsilon_t \tag{4.1}$$

where λ_i is the autoregressive coefficient corresponding to the i^{th} lag and ε_t is the residual noise at time t usually assumed uncorrelated Gaussian. The symmetric

Toeplitz covariance matrix for an AR(p) noise model can be described in terms of the Yule-Walker equations Yule (1927); Walker (1931) by solving these equations the AR coefficients for a given signal can be calculated iteratively. The AR noise model is a popular choice in the modelling of correlated noise in particular for fMRI due to attractive computational properties. In particular closed form expressions for the inverse of the covariance matrix for these models exists as described in Siddiqui (1958).

4.2 Linear Cross Correlation Analysis

We define the cross-correlation as a measure of the similarity between a potentially complex discrete function $\boldsymbol{f}$ and a reference function $\boldsymbol{g}$ at lag τ

$$(\boldsymbol{f} \star \boldsymbol{g})_\tau = \sum_j f_j^* g_{\tau+j}, \tag{4.2}$$

where * denotes the complex conjugate. Further we denote the cross-correlation between a signal $\boldsymbol{f}$ and itself the autocorrelation, e.g. $(\boldsymbol{f} \star \boldsymbol{f})_\tau$. In order to determine activity in fMRI studies the Pearson product-moment correlation coefficient estimate between the time series signal $\boldsymbol{y}$ and a reference time series $\boldsymbol{x}$ (at zero lag) has been suggested (Bandettini *et al.*, 1992, 1993)

$$r_{\boldsymbol{xy}} = \frac{\sum (x_i - \overline{\boldsymbol{x}})(y_i - \overline{\boldsymbol{y}})}{\sqrt{\sum (x_i - \overline{\boldsymbol{x}})^2 (y_i - \overline{\boldsymbol{y}})^2}}. \tag{4.3}$$

Here $\overline{x}$ denotes the sample mean of $\boldsymbol{x}$. The reference signal is typically based on some exceptions for the experiment, the correlation is then evaluation over the voxels in the brain and voxels where the correlation exceeds some threshold are declared active.

4.3 Statistical models and Bayes Theorem

A special case of what is now known as Bayes theorem was presented in an essay by Price (1763) with reference to his friend the, at the time of publication, deceased Reverend Thomas Bayes. Bayes theorem allows us to take into account prior information during modelling. In the following we will consider Bayes theorem used on a generative model $\mathscr{M}$. The generative model enables us to simulate an output vector which we will denote $\boldsymbol{y}$ given some parameters $\boldsymbol{\theta}$. Given a specific choice of noise model and parameters it is relatively easy to

calculate the probability of observing a specific output ($p(\boldsymbol{y}|\mathscr{M},\boldsymbol{\theta})$), we will denote this probability the likelihood. However, typically the quantities of interest are conditioned on the data and not the model. Bayes theorem

$$p(\mathscr{M},\boldsymbol{\theta}) = \frac{(\boldsymbol{y}|\mathscr{M},\boldsymbol{\theta})p(M,\theta)}{p(\boldsymbol{y})} \tag{4.4}$$

allows us to asses the probability of the model and parameters *after* observing a specific data vector $\boldsymbol{y}$ (posterior probability). The additional quantities that have entered in this equation are the probability of the model, $p(\mathscr{M},\boldsymbol{\theta})$, and the combined probability of observing the data over all possible models, $p(\boldsymbol{y})$. The probability of the model expresses the expectations about the model and parameters *prior* to observing the data (hence typically denoted the prior). The probability of the data in the denominator can be calculated by summing/integrating the likelihood over all possible models and parameters. Note that this quantity only depends on the data (constant given a specific observation vector) therefore its calculation is often omitted because it serves only as a normalisation factor (and due to the substantial computational complexity involved in its evaluation).

4.4 The General Linear Model

Linear models are very popular for the analysis of fMRI data (Friston *et al.*, 1994, 1995c; Boynton *et al.*, 1996) due their simple nature. We define the GLM

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{4.5}$$

Where $\boldsymbol{Y}$ denotes a $T \times N$ matrix containing the observed times series (of length T) as columns for each of the N voxels. $\boldsymbol{X}$ is a $T \times K$ matrix which we will denote the design matrix and $\boldsymbol{B}$ is a $K \times N$ parameter matrix (to be estimated). Finally, the $\boldsymbol{E}$ is the $T \times N$ residual matrix. To estimate values for the parameters we will proceed by choosing a probability density for the residual a popular and simple choice is the spatially independent matrix variate Normal distribution ($\boldsymbol{E} \sim \mathcal{M}_{T\times N}(\boldsymbol{0},\boldsymbol{I},\boldsymbol{\Sigma})$)

$$p(\boldsymbol{E}|\boldsymbol{\Sigma}) = (2\pi)^{\frac{-TN}{2}}|\boldsymbol{\Sigma}|^{-N/2}\exp\left(-\frac{1}{2}\operatorname{Tr}\left[\boldsymbol{E}^\top\boldsymbol{\Sigma}^{-1}\mathbf{E}\right]\right) \tag{4.6}$$

Note that the temporal covariance is assumed to be constant over space in the voxels included in the model. If we substitute $\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}$ for $\boldsymbol{E}$ in the expression above we can maximise the probability by differentiating with respect to the each of parameters and equating to zero. Hereby we can construct the following relation for the maximum likelihood estimate of the parameters $\boldsymbol{B}_{\mathrm{ML}}$

$$\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{B}_{\mathrm{ML}} = \boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \Rightarrow \boldsymbol{B}_{\mathrm{ML}} = (\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} \tag{4.7}$$

This estimate is also know as the generalised least squares (GLS) or pre-whitened maximum likelihood estimate. This can be motivated by studying the equivalent model where the data and design matrix respectively is transformed according to the relations $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}$ and $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$. In this equivalent the residual are now uncorrelated (scalar covariance). The ML for this pre-whitened problem is known as the ordinary least squares (OLS) solution. Even though the expressions above describe a multivariate model the estimation is essentially univariate due to the assumption of independence over voxels and we may therefore estimate any subset of voxels independently. We have further assumed that the design matrix is deterministic and that the covariance $\boldsymbol{\Sigma}$ is known. However, the main problem is obtaining a good estimate for the temporal covariance, one possible choice is to estimate it from the residuals $\overline{\boldsymbol{\Sigma}} = \frac{1}{1-K}\boldsymbol{E}^\top\boldsymbol{E}$ note that this estimate only has full rank and hence can be inverted if $N >= T + K$. The estimation of the temporal covariance is an ill-posed problem due to the many free parameters, hence the estimate of the covariance will be very dependent upon even small variations in the data. In order to reduce this problem and handle cases where $N < T + K$ a specific structural form or a regularised version of $\boldsymbol{\Sigma}$ is often used. By the use of an expectation-maximisation (EM) algorithm where the model parameters and the free parameters of the temporal covariance matrix are estimated alternatingly it is possible to take into account temporal correlations in the noise. An example of such an algorithm, where the temporal covariance matrix is assumed to be based on an AR(1) process is the restricted ML (reML) procedure applied in the SPM software package (Worsley and Friston, 1995). Note also that the temporal covariance need not be estimated from the same data that the model is specified for but may be based on based on an estimate over a larger region (Worsley, 2005). Techniques that rely on estimating priors from the data itself is typically known as empirical Bayesian techniques.

In the case where $\boldsymbol{X}^\top\boldsymbol{X}$ is singular we may still use the expressions given above by replacing the inverse by a pseudo inverse and K with the rank of $\boldsymbol{X}$. However, in this case the parameters of the model are not unique (in the sense that other parameters can produce the same residual) and interpretation of the parameters may be hampered. The ML estimates given above implicitly assumes that all parameter values are equally probable (flat parameter prior) in section 5.1 we will consider a case where prior expectations on the parameters can be included in the model.

4.5 Hypothesis Testing

In classical/Fisher statistics testing is usually done using the hypothesis testing framework. In this framework a null hypothesis is constructed reflecting how a

test statistic would behave if the effect of interest was not present. If the test statistic is sufficiently extreme as compared to the expected behaviour the data does not support the null hypothesis and we accept the alternative hypothesis. An example of this is the $\mathcal{T}$-test

$$t = \frac{\widehat{x} - \mu}{\widehat{\sigma}/\sqrt{n}} \tag{4.8}$$

Where x is the quantity to be tested, μ is the value of x that defines the null hypothesis, $\widehat{\sigma}$ is the sample standard deviation and n is the residual degrees of freedom (total number of samples minus the degrees of freedom used to estimate the model). Under the null hypothesis (no effect, Gaussian uncorrelated noise) the test statistic t is distributed according to a $\mathcal{T}$-distribution with n degrees of freedom. If x is based on two samples (two-sample $\mathcal{T}$-test/paired $\mathcal{T}$-test) the denominator can be calculated assuming either equal sample size and variances (polled-variance estimate), different sample sizes and equal variances, or different sample sizes and unequal variance through the Welch-Satterthwaite equation (Satterthwaite, 1946; Welch, 1947). Testing for effects in fMRI experiments is classically done by mass univariate hypothesis testing, where the effect of interest is tested on a voxel-wise basis. The voxel-wise $\mathcal{T}$-test in the GLM with pre-whitening reads

$$t_i = \frac{\boldsymbol{c}^\top \boldsymbol{b}_i}{\widehat{\sigma}_i(\boldsymbol{c}^\top \boldsymbol{b}_i)/\sqrt{n}}, \tag{4.9}$$

where the standard deviation estimate in the denominator is calculated according to $\widehat{\sigma}_i(\boldsymbol{c}^\top \boldsymbol{b}_i) = \sqrt{\widehat{\sigma}_i^2 \boldsymbol{c}^\top (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^\dagger \boldsymbol{c}}$. The contrast vector $\boldsymbol{c}$ defines a linear combination of effects (columns in $\boldsymbol{X}$) in the GLM to be tested. An alternative test useful for testing the significance of a combination of effects is the $\mathcal{F}$-test, here the ratio between the variance explained by two competing models (the null model and the alternative model) is tested. For a description of various tests and in particular the concept of contrasts see for example Christensen (1996).

The voxel-wise $\mathcal{T}$- and $\mathcal{F}$-test attempt to control the voxel-wise false positives (the probability of observing an effect when the null hypothesis is true is commonly referred to as an uncorrected p-value). However, when many voxels are considered this may not be desirable as even a quite low false positive rate can result in many non-active voxels being reported active. Under the assumption that all tests (voxels) are independent Bonferroni correction allow control of the so called family-wise error by simply dividing α by the number of (independent) tests. The family-wise error here refers to the probability of making one or more type-I errors. This correction, however, is typically considered too conservative due to the fact that voxels exhibit strong local dependence (especially after spatial smoothing). In order to take spatial dependence into account

while controlling the family-wise error rate corrections based on Gaussian random fields GRF have been suggested (Worsley *et al.*, 1996). Another alternative is to control the expected overall false discovery rate (FDR), i.e. the fraction of voxels declared active which on average can be expected to be false positives (Benjamini and Yekutieli, 2001; Benjamini and Hochberg, 1995; Genovese *et al.*, 2002).

The strong emphasis put on the null hypothesis unfortunately has a tendency to make hypothesis testing quite sensitive to the explicit assumptions made such as for example normality. For instance variance estimates are quite sensitive to outliers making assumption of Gaussian residuals/normality important in especially $\mathcal{F}$-tests.

The hypothesis testing framework is based on testing that the null hypothesis becomes sufficiently improbable, the Bayesian framework allows us to go beyond standard hypothesis testing in that we are able to assign probabilities to all the models of interest under the assumption that the prior distributions applied are appropriate.

4.5.1 Nuisance Variable Regression

In this section we describe ways to model nuisance effects through a linear model we will denote nuisance variable regression (NVR). We demonstrated the usefulness of this procedure in Lund *et al.* (2006) and showed that explicit modelling of various nuisance effects reduces violations of the normality and independence assumptions usually made in the GLM.

Noise related to physiological effects is a very prominent confounding factor in fMRI. The respiration cycle causes blood oxygenation changes, gross head movement and movement of organs in the abdomen (Raj *et al.*, 2001). Also the cardiac cycle introduces complicated artefacts such as inflow effects of magnetised blood and movement of the tissue near larger vessels that cannot be described by a simple rigid body transformation. In addition to increasing the signal variance these effects are known to introduce temporal correlations also known as non-white noise in the residual errors due to the oscillatory behaviour of these signals. This has implications in the analysis step as residual errors are usually assumed to be either independently identically distributed (i.i.d.) or to be described by a simple and sometimes global noise model such as an AR(1) process (Friston *et al.*, 2002). Removal of these physiological noise effects using high, low or band-pass filters are only possible if their frequency range is known and limited. Due to the long TR usually applied in fMRI to get whole brain coverage (typically 1-3 s) respiration and certainly the cardiac cycle is usually

under sampled causing aliasing. This undersampling causes the effect to alias into frequencies which could be both high and low within the sampling bandwidth rendering the physiological noise broad banded. Furthermore, broadening of the frequency range due to non-stationarity of physiological noise is strongly amplified by aliasing as we showed in Lund *et al.* (2006). Several attempts to model these effects using various techniques have been performed Glover *et al.* (2000); Hu *et al.* (1995); Lund *et al.* (2006); Thomas *et al.* (2002).

The NVR model (Lund *et al.*, 2006) includes three sets of regressors to account for known nuisance effects (β will refer to parameters to be determined from data for example by ML estimation):

1 A highpass filter in the form of a discrete cosine set (Worsley and Friston, 1995) to model low frequency drift caused by hardware instabilities:

$$N_L(t_n) = \beta_{L,1}\cos(\phi_L(t_n)) + \beta_{L,2}\cos(2\phi_L(t_n)) + \cdots + \beta_{L,p}\cos(p\phi_L(t_n)) \tag{4.10}$$

Where $\phi_L(t_n) = \pi/T$ and $p = \text{floor}(2T/T_L)$. T is the duration of the entire time-series, T_L is the period of the fastest oscillation to be removed, t_n is the time at which the first slice of volume n is sampled and floor denotes a function that returns the largest integer less than or equal to the argument.

2 A first order Volterra expansion of the movement parameters (Friston *et al.*, 1996) obtained from the rigid body re-alignment procedure to model residual movement effects still present after re-alignment:

$$\begin{aligned} N_M(t_n) =& \beta_{M,1}m_1(t_n) + \cdots + \beta_{M,6}m_6(t_n) \\ &+ \beta_{M,7}m_1(t_{n-1}) + \cdots + \beta_{M,12}m_6(t_{n-1}) \\ &+ \beta_{M,13}m_1^2(t_n) + \cdots + \beta_{M,18}m_6^2(t_n) \\ &+ \beta_{M,19}m_1^2(t_{n-1}) + \cdots + \beta_{M,24}m_6^2(t_{n-1}) \end{aligned} \tag{4.11}$$

Where $m_i(t_n)$ is the i^{th} rigid-body re-alignment parameter ($i = x, y, z, r_x, r_y, r_z$) corresponding to the volume acquired at time t_n.

3 Physiological noise regressors based on the retrospective image correction (RETROICOR) method described by Glover *et al.* (2000). This method models the physiological noise as a basis-set of sines and cosines representing the aliased frequencies of the cardiac and respiratory oscillations and their higher harmonics. Glover *et al.* (2000) used both linear and non-linear increasing phases for modelling of respiration effects. However, due to the comprehensive modelling of residual movement effects we here consider primarily to model the oxygenation-dependent part of the

respiration-induced noise, and thus only the phase increasing linearly with time is used. The RETROICOR regression variables using 5 harmonics of the cardiac and 3 harmonics of the respiratory signal then read:

$$\begin{aligned} N_P(t_n) =& \beta_{P,1}\sin(\phi_c(t_n)) + \beta_{P,2}\cos(\phi_c(t_n)) + \dots \\ &+ \beta_{P,9}\sin(5\phi_c(t_n)) + \beta_{P,10}\cos(5\phi_c(t_n)) \\ &+ \beta_{P,11}\sin(\phi_r(t_n)) + \beta_{P,12}\cos(\phi_r(t_n)) + \dots \\ &+ \beta_{P,15}\sin(3\phi_r(t_n)) + \beta_{P,16}\cos(3\phi_r(t_n)) \end{aligned} \tag{4.12}$$

Where $\phi_c(t_n)$ and $\phi_r(t_n)$ are the phases cardiac and respiratory phases respectively assigned to the volume acquired at time t_n. The phase here refer to the temporal distance from the first slice in the current volume to the last peak in the cardiac or respiratory reference time-course, divided with the peak to peak interval at that time-point. Here it is assumed that the pulse and respiration rate stays constant within the acquisition of one volume, however, the model is still capable of capturing non-stationary in the cardiac and respiration time series. The acquisition time of volume n, t_n, is considered the time at which the middle slice (time-wise) is acquired.

4.6 Unsupervised Analysis

4.6.1 Factor Analysis

The factor analysis (FA) model developed by Spearman (1904) describes how matrix-variate data can be described by a linear mixing of factors (sometimes referred to as sources)

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} + \boldsymbol{E}. \tag{4.13}$$

Where $\boldsymbol{Y}$ is a data matrix (for example time $\times$ voxels), the rows of $\boldsymbol{S}$ forms the factors (spatial) and the columns of the mixing matrix, $\boldsymbol{A}$, contains the strength over time and weight for each of the factors as the rows. $\boldsymbol{E}$ is the residual error (typically assumed to be Gaussian with diagonal covariance). Note that the model is identical to the GLM. Here, however, we aim to infer both the mixing matrix and the factors. Unfortunately, the model above has a major flaw in terms of its uniqueness, this can be seen by considering an alternative factorisation

$$\boldsymbol{Y} = (\boldsymbol{A}\boldsymbol{Q}^{-1})(\boldsymbol{Q}\boldsymbol{S}) + \boldsymbol{E}.$$

Where $\boldsymbol{Q}$ is any invertible matrix, this alternative model is clearly equivalent to the former model in the sense that it provides the exact same residual error. For $\boldsymbol{A}$ and $\boldsymbol{S}$ to be uniquely defined it is necessary to impose additional constraints in the model.

Resolving the ambiguity of the FA decomposition has been the topic of much research and several different methods have been suggested a few of which are described in the following sections.

4.6.2 Principal Component Analysis

The principal component analysis (PCA) also known as the Karhunen-Loève transform dates at least back to Pearson (1901) and has since been applied to a wide range of problems in many scientific fields. The PCA procedure mounts to consecutively finding the directions in the data accounting for the most variance. These directions define the principal components here encoded as the rows of the matrix $\boldsymbol{U}$

$$\boldsymbol{U}^\top \boldsymbol{Y} = \boldsymbol{\Sigma} \boldsymbol{V}^\top \tag{4.14}$$

We here without loss of generality assume that the data matrix has zero empirical mean (e.g. the empirical mean has been subtracted). Also we require that $\boldsymbol{\Sigma}$ is a diagonal matrix and that both $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal (e.g. $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$). Typically, the principal components are sorted according to the amount of variance of the data they explain. Hence the diagonal of $\boldsymbol{\Sigma}$ is non increasing. This model is a severely constrained version of the FA model, the advantage of which is that it is uniquely defined under the assumption that all eigenvalues of the sample covariance matrix ($\boldsymbol{Y}\boldsymbol{Y}^\top$) are distinct.

The PCA decomposition can be performed efficiently using the singular value decomposition (SVD where the cost function $\|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{S}\|_F^2$ is minimised subject to the constraints $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{\Sigma}$ being diagonal (Golub and Van Loan, 1996; Golub and Kahan, 1965).

Even though orthogonality causes PCA to be unique this constraint is normally not meaningful for real data. Nevertheless approximate PCA decompositions (retaining only the components corresponding the largest singular values) are useful for identifying the signal subspace. Such decompositions are often used for dimensionality reduction (compression) or for forming a good starting point for other methods such as ICA.

4.6.3 Independent Component Analysis

The concept of independent component analysis (ICA) was introduced in a paper by Comon (1994). The main concept is to resolve ambiguity of FA type decompositions by the assumption of independence of the sources, which is a

good assumption for many data types. For fMRI data ICA has become a popular explorative analysis tool and has been used to identify both signal and noise in the data (McKeown *et al.*, 1998b,a; McKeown and Sejnowski, 1998; McKeown *et al.*, 2003; Thomas *et al.*, 2002; Beckmann and Smith, 2004). We begin by writing the noiseless ICA (or FA) model

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{S} \tag{4.15}$$

We here motivate the model as a blind source separation problem (BSS), i.e. we interpret the signal as a noise free mixing of the latent sources encoded in the rows of $\boldsymbol{S}$. We now aim to estimate the un-mixing matrix $\boldsymbol{W} = \boldsymbol{A}^{-1}$ that will reconstruct the sources as $\boldsymbol{S} = \boldsymbol{W}\boldsymbol{X}$. We note that mixtures of independent signals will not be independent and tends to become more Gaussian than the signals themselves due to the central limit theorem. Hence, by estimating $\boldsymbol{W}$ such that the rows of $\boldsymbol{S}$ become as non-Gaussian as possible while assuming that the rows are independent should allow us to recover the original sources (up to permutation and scaling). We note that this approach is only feasible if the sources are not Gaussian (with the exception of one which can be allowed to be Gaussian)(Hyvärinen, 1999).

We can write the joint density of the signal (given that we know the correct $\boldsymbol{W}$, which we assume to be deterministic) as

$$p(\boldsymbol{X}|\boldsymbol{W}) = \prod_i p_s(\boldsymbol{w}_{:,i}\boldsymbol{X})|\boldsymbol{W}|, \tag{4.16}$$

where the $|\boldsymbol{W}|$ is found as the determinant of the Jacobian of $\boldsymbol{S}$ with respect to $\boldsymbol{X}$. For maximisation purposes, i.e. ML ICA, we will consider the log-likelihood

$$\log p(\boldsymbol{X}|\boldsymbol{W}) = \sum_{ij} \log p_s(\boldsymbol{w}_{:,i}\boldsymbol{x}_j) + J \log |\boldsymbol{W}|. \tag{4.17}$$

Here, the source distribution should be selected in accordance with prior expectations an example is the Laplace prior which would correspond to determining the sources such that the l_1-norm is minimal. The determinant of the un-mixing matrix ensures that the un-mixing/mixing matrix is invertible. This approach of ML ICA is equivalent to the information maximisation approach given in (Bell and Sejnowski, 1995).

Other methods for ICA include diagonalisation of higher order moments, cumulants (Comon, 1994) or autocorrelation functions (Molgedey and Schuster, 1994) of the source matrix. Procedures such as the varimax rotation criterion (Carroll, 1953; Kaiser, 1958) and related methods developed to resolve the ambiguity of FA models can also be considered ICA algorithms, however, these were not motivated directly through independence of the sources.

4.6.4 Sparse Coding

We once again consider the FA model and proceed by construction of a combined likelihood and prior of the model. The likelihood term, $p_\varepsilon(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{S})$, favours data reconstruction and the prior term, $p_s(\boldsymbol{S})$, favours sparsity of the sources.

$$p_{sc} = p(\boldsymbol{Y},\boldsymbol{A},\boldsymbol{S})p_s(\boldsymbol{S}) \tag{4.18}$$

The reconstruction term is often taken to be based on Gaussian noise, i.e. based on the sum of squared error ($\|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{S}\|_F^2$), i.e. the l_2-norm of the reconstruction. The number of non-zero elements in $\boldsymbol{S}$, i.e. the l_0-norm appears to be an appealing way to measure sparsity of $\boldsymbol{S}$. Unfortunately, ML estimation using priors based on this quantity leads to N-P hard optimisation problems and in practice the densities based on for example the Laplace prior (l_1-norm) is often chosen serving as a convex proxy for the l_0-norm. By taking the logarithm, switching sign and dropping normalisation constants in equation 4.18 we arrive at the objective

$$C_{SC} = \|\boldsymbol{Y}-\boldsymbol{A}\boldsymbol{S}\|_F^2 + \lambda\|\boldsymbol{S}\|_1, \tag{4.19}$$

where λ serves as a sparsity regularisation strength. The minimisation of the objective is often approached by a Gauss-Siedel strategy, i.e. alternating between solving $\boldsymbol{S}$ for fixed $\boldsymbol{A}$ and vice versa. Normally, the norm of the components in $\boldsymbol{A}$ is kept fixed by applying a Lagrange multiplier approach (Lee *et al.*, 2007) or by projected interior point methods (Olshausen and Field, 1996). Solving $\boldsymbol{S}$ for fixed $\boldsymbol{A}$ based on the objective in equation 4.19 is the well known Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) or basis pursuit de-noising (BPD) problem (Shaobing and Donoho, 1994). This convex optimisation problem has no known closed form solution and iterative methods are typically applied. One such example is the Least Angle Regression and Selection (LARS) algorithm(Efron *et al.*, 2004) described in appendix A.2.4 or the equivalent homotropy (Osborne *et al.*, 2000) method which allows minimisation of this objective function for all values of λ at the cost of an OLS solution. Other approaches include conjugate gradient (Olshausen and Field, 1996), turning the problem into a non-negative quadratic programming problem with negativity constraints of doubled size as done in the SPARSELAB (`http://www.sparselab.stanford.edu`) or using an iterative re-weighted least squares (IRLS) procedure (Blumensath and Davies, 2004). Yet another simple alternative is the modified gradient method given in appendix A.2.1 where the update is spilt into two to avoid oscillations around zero.

One way to approach the problem of selecting an appropriate value of λ would be to consider prior expectations (for example expected variance of $\boldsymbol{S}$). In practice, however, several different sparsity penalties are applied and the one providing the best tradeoff between sparsity and reconstruction is selected. One possibility

in the selection of this optimal tradeoff is large sample approximations such as the Bayesian Information Criterion (BIC) or Akaike's Information Criterion (AIC). In practice, however, somewhat more heuristic measures such as the L-curve (identifying the region of maximum curvature in a plot of for example the reconstruction error as a function of the l_0-norm or l_1-norm) yields better results.

A related model is the sparse PCA model suggested in (Zou *et al.*, 2006)

$$\boldsymbol{Y} \approx \boldsymbol{AWY},$$

where $\boldsymbol{W}$ is regularised with a sparsity prior. The rationale behind this model is to be able to explain data as a linear combination as few of the original data points as possible (thus sources are $\boldsymbol{S} = \boldsymbol{WY}$). This model can be estimated through the minimisation of the following objective

$$\begin{array}{cc} \text{argmin}_{\boldsymbol{A},\boldsymbol{W}} & \|\boldsymbol{Y} - \boldsymbol{AWY}\|_F^2 + \lambda_1\|\boldsymbol{W}\|_1 + \lambda_2\|\boldsymbol{W}\|_F^2 \\ \text{subj. to} & \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{I} \end{array}.$$

Where λ_1 and λ_2 are regularisation parameters for the l_1-norm and l_2-norm respectively. Note that $\boldsymbol{A}$ is restricted orthonormal, this together with a small l_2-norm regularisation ensures that the model is equivalent to standard PCA when no sparsity regularisation is chosen ($\lambda_1 = 0$). We demonstrated this model on fMRI data in Sjöstrand *et al.* (2006).

4.6.5 Non-negative Matrix Factorisation

The non-negative matrix factorisation model (also known as positive matrix factorisation) as described by Paatero and Tapper (1994) and later reinvented by Lee and Seung (1999) is a version of the FA model where the model parameters are constrained positive. Due to the non-negativity constraint cancellation effects do not occur providing an often desired part-based representation. The model has proved useful for a wide range of data such as text, PET (Lee *et al.*, 2001) and spectral data (Gobinet *et al.*, 2004). Decomposition of unconstrained data may be achieved by considering the amplitude of a spectral representation (Smaragdis and Brown, 2003) or simply adding a constant offset to the data. The non-negative matrix factorisation model can be estimated using the popular multiplicative updates (Lee and Seung, 1999) or over relaxed bound optimisation (Salakhutdinov *et al.*, 2003) described in appendices A.2.2 and A.2.3 respectively. Other approaches include classical non-negative quadratic programming using active set procedures (Lawson and Hanson, 1974; Bro and de Jong, 1997) or methods that rely on projecting the gradient to the positive orthant as described in Lin (2007). Unfortunately, non-negativity constraints

does not in general alleviate the problem of non-uniqueness (Donoho and Stodden, 2004) and often additional constraints such as sparseness (Donoho and Stodden, 2004) is imposed. In Mørup *et al.* (2008a) we presented a sparse version of the NMF optimised using a non-negative version of the LARS algorithm (see appendix A.2.5).

4.7 Multi-way Decompositions

4.7.1 Canonical Decomposition / Parallel Factor Analysis

The parallel factor analysis (PARAFAC) also named canonical decomposition (CANDECOMP) as suggested by (Carroll and Chang, 1970; Harshman, 1970) is a simple generalisation of the FA decomposition model to n-way tensors. This model, which we will denote the CANDECOMP/PARAFAC (CP) model, states how a n-way tensor is decomposed into a sum of rank one components

$$y_{i_1,i_2,\ldots,i_N} \approx \sum_{d=1}^{D} a^{(1)}_{i_1,d} a^{(2)}_{i_2,d} \ldots a^{(N)}_{i_n,d}. \tag{4.20}$$

The approximation will be exact when D is higher than or equal to the rank of the tensor $\mathcal{Y}$ (which defines the tensor rank). The 3-way version of this model is illustrated graphically in figure 4.1. In fMRI one example of data structured in a 3-way tensor is data measured over multiple subjects, here the modes are; space, time and subjects. Due to the very restrictive nature of the CP decomposition it is contrary to the FA model unique under mild conditions. In a rigorous proof by Kruskal (1976) several sufficient conditions for the decomposition to be unique is given. The most popular of these generalised to n-way (Sidiropoulos and Bro, 2000) states that decomposition is unique when

$$\sum_{n=1}^{N} k_{\boldsymbol{A}^{(n)}} \geq 2D + N - 1, \tag{4.21}$$

here $k_{\boldsymbol{A}^{(n)}}$ denotes the k-rank (Kruskal rank) of the matrix $\boldsymbol{A}^{(n)}$ which is the maximal number of rows that can be selected from $\boldsymbol{A}^{(n)}$ such that all these possible sets are linearly independent (the maximum independent set). Note that it follows from this criterion that the CP decomposition may be unique even when the number of components are greater than the rank of every mode in the tensor.

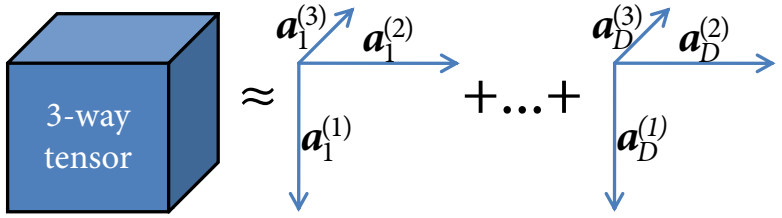


Figure 4.1: **The CP model.** The figure illustrates how a 3-way tensor is approximated by a sum of rank-1 components in the 3-way CP model.

4.7.2 Estimating the CP model

Unfortunately, the CP model has no known closed form solution identifying the correct subspace as was the case with the SVD for the FA model when assuming independent Gaussian noise with scalar covariance. However, by applying the matricising operation and the Khatri-Rao product the model can be restated such that the estimation for each mode becomes equivalent to a regular FA model

$$\boldsymbol{Y}_{(n)} \approx \boldsymbol{A}^{(n)}\boldsymbol{Z}, \tag{4.22}$$

where $\boldsymbol{Y}_{(n)}$ denotes the matrix obtained by matricising $\mathcal{Y}$ over the n'th mode and $\boldsymbol{Z}^\top$ is the Khatri-Rao product over the remaining modes,

$$\boldsymbol{Z}^\top = \boldsymbol{A}^{(N)} \odot \boldsymbol{A}^{(N-1)} \odot \ldots \odot \boldsymbol{A}^{(n+1)} \odot \boldsymbol{A}^{(n-1)} \odot \ldots \odot \boldsymbol{A}^{(1)}.$$

By solving alternating between each of the modes the alternating least squares (ALS) algorithm is formed. Note that if we assume independent Gaussian noise with scalar covariance there is a closed form solution for each mode

$$\boldsymbol{A}^{(n)} \leftarrow \boldsymbol{Y}_{(n)}\boldsymbol{Z}\boldsymbol{H}^{-1},$$

where

$$\boldsymbol{H} = \boldsymbol{A}^{(N)\top}\boldsymbol{A}^{(N)} \bullet \ldots \bullet \boldsymbol{A}^{(n+1)\top}\boldsymbol{A}^{(n+1)} \bullet \boldsymbol{A}^{(n-1)\top}\boldsymbol{A}^{(n-1)} \bullet \ldots \bullet \boldsymbol{A}^{(1)\top}\boldsymbol{A}^{(1)}.$$

Note that the matrix inversion is on the often small factors $\times$ factors matrix. Unfortunately evaluating the Khatri-Rao product to form $\boldsymbol{Z}$ is computationally

intensive for large scale problems, a much more computationally efficient algorithm can be implemented by pre-multiplying the largest mode with the data before solving for the remaining modes (Tomasi, 2006).

4.7.3 Tucker

The Tucker model (Tucker, 1966) can be written as

$$y_{i_1,i_2,\ldots,i_N} \approx \sum_{j_1,j_2,\ldots,j_N} g_{j_1,j_2,\ldots,j_N} a^{(1)}_{i_1,j_1} a^{(2)}_{i_2,j_2} \ldots a^{(N)}_{i_N,j_N}, \tag{4.23}$$

and can be considered an extended version of the CP where the core array $\mathcal{G}$ serves to model linear interactions between the modes. By utilising the n-mode matricising operation and kronecker product we can restate the Tucker model as a sequence of regular FA problems

$$\boldsymbol{Y}_{(n)} \approx \boldsymbol{A}^{(n)} \boldsymbol{Z}, \tag{4.24}$$

where

$$\boldsymbol{Z} = \boldsymbol{G}_{(n)} \left(\boldsymbol{A}^{(N)} \otimes \boldsymbol{A}^{(N-1)} \otimes \ldots \otimes \boldsymbol{A}^{(n+1)} \otimes \boldsymbol{A}^{(n-1)} \otimes \ldots \otimes \boldsymbol{A}^{(1)} \right)^{\top}.$$

As a result estimation for each mode can be performed by using pseudo-inverses for the least squares objective (Andersson and Bro, 1998). Estimation of the core array can be performed utilising the n-mode tensor product, $\times_n$, and regular matrix pseudo-inverse

$$\mathcal{G} = \mathcal{Y} \times_1 \boldsymbol{A}^{(1)\dagger} \times_2 \boldsymbol{A}^{(2)\dagger} \times_3 \ldots \times_N \boldsymbol{A}^{(N)\dagger}. \tag{4.25}$$

Note that the Tucker model is equivalent to the CP model when the core tensor is diagonal. By estimating the core tensor corresponding to an estimated CP model and measuring the 'deviation' from a diagonal tensor the core consistency diagnostic criterion is formed (Bro and Kiers, 2003). By examining when there is a strong drop in core consistency (when adding more factors) it is possible to asses how many factors are appropriate for describing the data at hand.

Chapter 5

New Models

> Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
>
> *George E. P. Box (18 October 1920 –)*
> *Box and Draper, Empirical Model-Building, p. 74, Wiley*

5.1 Poor-mans Bayesian General Linear Model

This section describe how using the conjugate prior allows efficient computation of the model evidence/Bayes factor in a linear model serving as a computationally 'inexpensive' alternative to iterative methods such as numerical integration or variational Bayes.

The description of the GLM given in section 4.4 relied on ML for estimation of the parameters (flat prior) and with Gaussian assumptions on the residual error term. The Bayesian framework described in section 4.3 allows us to incorporate prior information in the model by the use of a meaningful prior distribution on the model parameters. Typically, this process involve the use of methods such as variational Bayes or numerical integration to obtain the posterior distribution Penny *et al.* (2003); Kershaw *et al.* (1999). However, if many models are to be compared solutions based on these iterative methods may be intractable. We

therefore suggest the use of conjugate priors in which case the posterior distribution and Bayes factor can be obtained as a closed form expression. Earlier work Hansen *et al.* (2002) applied a Normal-Inverse-Gamma (NIG) conjugate prior for model selection. In this case noise is assumed Gaussian identically and independently distributed (i.i.d., 'white noise'). It is, however, well established that noise in fMRI exhibits temporal correlations, see e.g., (Weisskoff *et al.*, 1993) which is typically dealt with by a pre-whitening procedure assuming a known (typically estimated from data) temporal covariance structure.

Here we aim to derive a detection framework based on the natural conjugate prior for the GLM allowing for arbitrarily correlated normal noise. For this we apply the normal-inverse Wishart (NIW) prior. The choice of a conjugated prior carries with it some constraints, in particular, it implies a relation between the noise covariance and model parameter prior. Such a constraint may or may not be relevant for a given setup. However, they lead to much more computationally efficient estimates, since the relevant Bayes factors are closed form expressions. It is well established that integration over model parameters can be carried out analytically choosing a Gaussian prior. However, here we consider also the integral over all possible realisations of the noise structure (i.e. average over all possible correlated noise model). This should improve stability in cases where the exact noise covariance is not known. Some generalisations of the natural conjugate prior exists for the GLM. However, these do not lead to closed form expressions for the Bayes factor (Press, 1982; Daz-Garca and Ramos-Quiroga, 2003). We proceed by deriving a closed form expression for Bayes factors in the GLM using the NIW prior, then we consider assigning meaningful hyper-parameters to the prior distribution. Given the choice of prior parameters we give expressions for the posterior parameter distribution and the posterior SNR. Finally, we consider some extensions of the framework to allow for adaptive regularisation using empirical Bayes and discuss the possibility of estimating hyper-parameters for the noise covariance prior.

In the following we will consider the linear model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is a data vector of length T, $\boldsymbol{X}$ is a design matrix consisting of K known components, $\boldsymbol{\beta}$ is an unknown parameter vector of length K and ε is the residual error assumed to be normal distributed with zero mean and unknown covariance ($\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$). With these assumptions the model likelihood is given by

$$p(\boldsymbol{y}|\mathcal{M}) = p(y|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = (2\pi)^{-T/2} \exp\left[-\tfrac{1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^{\top} \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right]. \tag{5.1}$$

Where $|\boldsymbol{\Sigma}|$ denotes the determinant of the matrix $\boldsymbol{\Sigma}$. In the following section we will define prior distributions for the model parameters $\boldsymbol{\beta}$ and the noise covariance $\boldsymbol{\Sigma}$.

Prior distribution

During the derivation we will consider a design matrix that is extended to have full rank (eg. a design matrix $\boldsymbol{X}_f$ with $T-K$ columns added). Note that we will eliminate this restriction later by forcing the parameters of the additional parameters to zero choosing a prior with zero variance on these parameters. We introduce the reconstructed signal $\boldsymbol{z} = \boldsymbol{X}_f\boldsymbol{\beta}$ and rewrite the model likelihood

$$p(y|\boldsymbol{z},\boldsymbol{\Sigma}) = (2\pi)^{-T/2}\exp\left[-\tfrac{1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{z})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{z})\right]. \tag{5.2}$$

We introduce the multivariate normal prior for the reconstructed signal $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Sigma})$

$$p(\boldsymbol{z}|\boldsymbol{Q},\boldsymbol{\Sigma}) = (2\pi)^{-T/2}|\boldsymbol{Q}|^{\frac{1}{2}}\exp\left[-\tfrac{1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\boldsymbol{z}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{z}\right]. \tag{5.3}$$

Notice that we have defined the prior for the reconstructed signal instead of the model parameters as it is typically done (Goutte *et al.*, 2000; Woolrich *et al.*, 2004). However, in this case where we have a Normal distributed prior and a linear relation between the model and the reconstructed signal the two has a simple relation (eg. if the parameters have covariance $\boldsymbol{V}$ the reconstructed signal will have covariance $\boldsymbol{X}_f\boldsymbol{V}\boldsymbol{X}_f^\top$). Here we will not require the product of $\boldsymbol{Q}^{-1}$ and $\boldsymbol{\Sigma}$ to be symmetric, however, we note that in cases where an asymmetric part exists this part will not contribute when we later integrate over this distribution.

Further, we will apply the inverted Wishart distribution for the noise covariance $\boldsymbol{\Sigma} \sim \mathcal{W}_m^{-1}(\boldsymbol{B})$

$$p(\boldsymbol{\Sigma}|\boldsymbol{B},m) = (2)^{-\frac{mT}{2}}|\boldsymbol{B}|^{\frac{m}{2}}\,\Gamma_T(\tfrac{m}{2})^{-1}\exp\left[-\tfrac{m+T+1}{2}\log|\boldsymbol{\Sigma}| - \tfrac{1}{2}\operatorname{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right)\right], \tag{5.4}$$

where the $\Gamma_T(a)$ is the multivariate gamma function given by

$$\Gamma_T(a) = \int_{\boldsymbol{G}\in\Omega}\exp(-\operatorname{Tr}\boldsymbol{G})\,|\boldsymbol{G}|^{a-\frac{(T+1)}{2}}\,d\boldsymbol{G} = \pi^{\frac{T(T-1)}{4}}\prod_{j=1}^{T}\Gamma(a+\tfrac{(j-1)}{2}),$$

where $\boldsymbol{\Omega}$ is the complete set of all $(T \times T)$ positive definite matrices and $\Gamma(a)$ denotes the ordinary Gamma function. Note the specific parametrisation of the inverted Wishart distribution; the noise covariance scale matrix is denoted $\boldsymbol{B}$ while m is the degrees of freedom. By multiplying the two prior distributions we obtain the NIW joint prior distribution (see appendix A.1.5) given the three hyper-parameters $\boldsymbol{Q}, \boldsymbol{B}$ and m

$$\begin{aligned} p(\boldsymbol{z}, \boldsymbol{\Sigma} | \boldsymbol{Q}, \boldsymbol{B}, m) =& (2)^{-\frac{(m+1)T}{2}} (\pi)^{-\frac{T}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} |\boldsymbol{B}|^{\frac{m}{2}} \Gamma_T(\tfrac{m}{2})^{-1} \\ & \times \exp\left[-\tfrac{m+T+2}{2} \log |\boldsymbol{\Sigma}|\right] \\ & \times \exp\left[-\tfrac{1}{2} \boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q} \boldsymbol{z}\right] \\ & \times \exp\left[-\tfrac{1}{2} \operatorname{Tr}\left(\boldsymbol{B} \boldsymbol{\Sigma}^{-1}\right)\right]. \end{aligned} \tag{5.5}$$

Posterior distribution

According to equation 4.4 the posterior distribution is proportional to the product of the likelihood and the prior distribution. We form this joint model and prior distribution as the product of equations 5.2 and 5.5

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\Sigma} | \boldsymbol{Q}, \boldsymbol{B}, m) =& (2)^{-\frac{(m+2)T}{2}} (\pi)^{-T} |\boldsymbol{Q}|^{\frac{1}{2}} |\boldsymbol{B}|^{\frac{m}{2}} \Gamma_T(\tfrac{m}{2})^{-1} \\ & \times \exp\left[-\tfrac{m_p+T+2}{2} \log |\boldsymbol{\Sigma}|\right] \\ & \times \exp\left[-\tfrac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_z)^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}_p (\boldsymbol{z} - \boldsymbol{\mu}_z)\right] \\ & \times \exp\left[-\operatorname{Tr}\left(\tfrac{1}{2} \boldsymbol{B}_p \boldsymbol{\Sigma}^{-1}\right)\right]. \end{aligned} \tag{5.6}$$

For a more detailed derivation see Madsen and Hansen (2007). The resulting distribution given above is by conjugacy again of the NIW type with updated hyper-parameters ($\boldsymbol{Q}_p$, $\boldsymbol{B}_p$, m_p and an additional parameter $\boldsymbol{\mu_z}$ the reconstructed mean which we assumed to be zero in the prior distribution). The update parameters are given by

$$\begin{aligned} \boldsymbol{Q}_p =& \boldsymbol{I}_T + \boldsymbol{Q}, \\ \boldsymbol{B}_p =& \boldsymbol{B} + (\boldsymbol{I}_T + \boldsymbol{Q}^{-1})^{-1} \boldsymbol{y} \boldsymbol{y}^\top, \\ m_p =& m + 1, \\ \boldsymbol{\mu_z} =& (\boldsymbol{I}_T + \boldsymbol{Q})^{-1} \boldsymbol{y}. \end{aligned}$$

The expression for the posterior distribution in equation 5.6 depends on the model parameters $\boldsymbol{\beta}$ and the noise covariance $\boldsymbol{\Sigma}$ in order to remove the dependence upon these parameters we proceed by marginalising the distribution with respect to these. This marginalisation yields simply the NIW normalisation integral (see appendix A.1.5) and we obtain

$$p(\mathcal{M}|\boldsymbol{y}, \boldsymbol{Q}, \boldsymbol{B}, m) \propto \frac{|\boldsymbol{B}|^{\frac{m}{2}} \left|\boldsymbol{I}_T + \boldsymbol{Q}^{-1}\right|^{\frac{-1}{2}}}{\left|\boldsymbol{B} + \left(\boldsymbol{I}_T + \boldsymbol{Q}^{-1}\right)^{-1} \boldsymbol{y}\boldsymbol{y}^\top\right|^{\frac{m+1}{2}}}. \tag{5.7}$$

This expression was derived under the assumption that the design matrix has full rank. However, by forcing the parameter estimates of the last $T-K$ columns in the design matrix to zero (by having zero prior variance for these) and considering assignment of reasonable prior hyper-parameters the following expression can be found (see (Madsen and Hansen, 2007) for further details)

$$p(\mathcal{M}|\boldsymbol{y}, v, \boldsymbol{B}, m) \propto \frac{b^{-\frac{T}{2}} |\boldsymbol{B}_{prior}|^{-\frac{1}{2}} |\boldsymbol{A}|^{\frac{1}{2}}}{\left(1 + \frac{1}{b}\left(\boldsymbol{y}^\top \boldsymbol{B}_{prior}^{-1} \boldsymbol{y} - v\boldsymbol{y}^\top \boldsymbol{B}_{prior}^{-1} \boldsymbol{X} \boldsymbol{A} \boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1} \boldsymbol{y}\right)\right)^{\frac{m+1}{2}}}, \tag{5.8}$$

where $b = \frac{(m-T-1)\boldsymbol{y}^\top \boldsymbol{y}}{\mathrm{Tr}(\boldsymbol{B}_{prior})}$, $v = \frac{\mathrm{Tr}(\boldsymbol{B}_{prior})}{(m-T+1)\,\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X})}$, $\boldsymbol{A} = \left(\boldsymbol{I}_K + v\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1} \boldsymbol{X}\right)^{-1}$ and $\boldsymbol{B}_{prior}$ is a hyper-parameter matrix set according to the expected covariance structure of the noise.

In the case where $\boldsymbol{B}_{prior}$ is chosen to be the identity matrix the expression given in equation (5.8) simplifies to the expression obtained with the white noise Normal-Inverse Gamma (NIG) prior (Hansen *et al.*, 2002), however, with modified interpretations of the hyper-parameters. Note that we here integrate over all possible realisations of correlated noise according to the Wishart prior.

Parameter estimates

The posterior reconstructed signal will follow a multivariate $\mathcal{T}$-distribution with $m - T + 2$ degrees of freedom, covariance matrix $\boldsymbol{Q}_p^{-1} \boldsymbol{B}_p$ and mean $\boldsymbol{\mu}_{\boldsymbol{z}}$. Here the posterior mode (MAP) estimate is simply the mean $\boldsymbol{\mu}_{\boldsymbol{z}}$ and we can obtain

the posterior parameter estimates ($\boldsymbol{\beta}_{MAP}$) directly by simply dropping the multiplication with $\boldsymbol{X}$ in the expression for $\boldsymbol{\mu_z}$

$$\begin{aligned}\boldsymbol{\mu}_z &= (\boldsymbol{I}_T + \boldsymbol{Q})^{-1}\boldsymbol{y} = \boldsymbol{Q}^{-1}(\boldsymbol{I}_T + \boldsymbol{Q}^{-1})^{-1}\boldsymbol{y} \\ &= v\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}(\boldsymbol{I}_T + v\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1})^{-1}\boldsymbol{y} \\ &= v\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{y} \Rightarrow \\ \boldsymbol{\beta}_{MAP} &= v\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{y}.\end{aligned} \tag{5.9}$$

In the limit where $v \to \infty$ this estimate coincides with the standard pre-whitened ML likelihood estimate.

$$\boldsymbol{\beta}_{ML} = (\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{y}. \tag{5.10}$$

Posterior signal to noise ratio

The posterior distribution contains information about both the detected signal and the noise. We will therefore derive a quantity that we denote the posterior SNR which can provide information about how heavily the signal was contaminated by noise. We estimate the posterior SNR as the expected mean of the reconstructed signal squared divided by the trace of the noise variance

$$\begin{aligned}SNR_{posterior} &= \frac{\boldsymbol{\mu_z}^\top \boldsymbol{\mu_z}}{\mathrm{Tr}(\frac{1}{m-T}\boldsymbol{B}_p)} \\ &= \frac{v^2\boldsymbol{y}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{y}}{\boldsymbol{y}^\top\boldsymbol{y} - \frac{1}{m-T}(v\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{B}_{prior}^{-1}\boldsymbol{y})}.\end{aligned} \tag{5.11}$$

We can now evaluate a bound for the posterior SNR which we will use to set a reasonable value for the prior hyper-parameter m. In the limit where the input signal $\boldsymbol{y}$ is contained within the subspace defined by the design matrix $\boldsymbol{X}$ and $v \to \infty$ (fully reconstructed signal) the posterior SNR (equation 5.11) will approach the value $\frac{m-T}{m-T-1}$, hence this is an upper bound for the posterior SNR. The prior signal covariance is finite when $m > T+1$. Choosing the lowest possible integer $m = T+2$ limits the posterior SNR to two and therefore this does not constitute a weak prior. In the case of maximum likelihood estimation the SNR scales with $\sqrt{T}$. Here, we can get a similar behaviour by setting $m = T+1+1/\sqrt{T}$ making the maximum obtainable (for $v \to \infty$) posterior SNR ratio

scale roughly with the square root of the input data dimension. In this relation it should be noted that m has no influence on the MAP parameter estimate.

In figure 5.1 we use the proposed detection framework to detect the correct noise model (based on NVR regressors) for each voxel in a visual experiment. Results for detection of visual activation and the estimated posterior SNR is displayed in figures 5.2 and 5.3 respectively. For more details, simulations and discussion see appendix B.1.

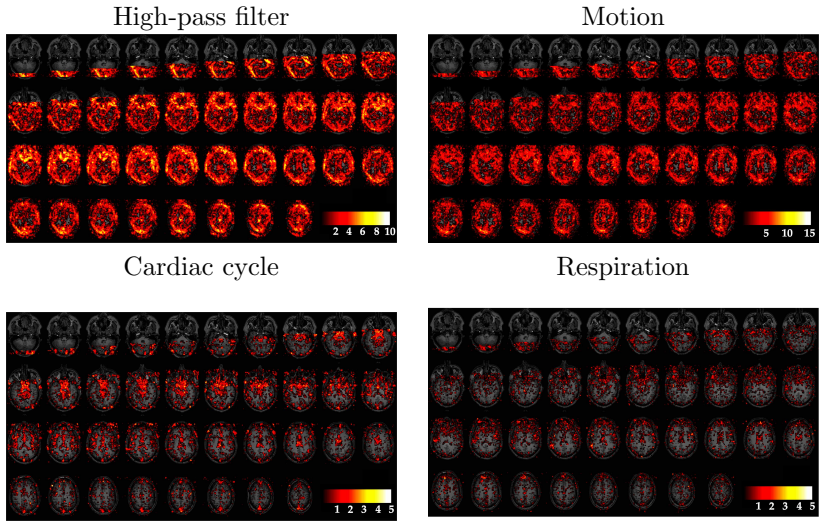


Figure 5.1: **Nuisance effects.** The figures show the detected nuisance model complexity overlayed on the anatomical volume for each of the nuisance effects indicated by the headings. Note that the colouring relates to the detected model complexity and not directly to the significance of the effects. Effect of the high-pass filter is widely distributed over the brain whereas motion is most prominent at the edges of the brain. At the large arteries of the brain such as the Circle of Willis and the Medial Cerebral Artery the detected model complexity is highest as it is to be expected. Respiratory effects are present at the edges of the brain as well as in the ventricles. The colour bars indicate either the model complexity K (high-pass filter and motion) or the order of the expansion (cardiac cycle and respiration).

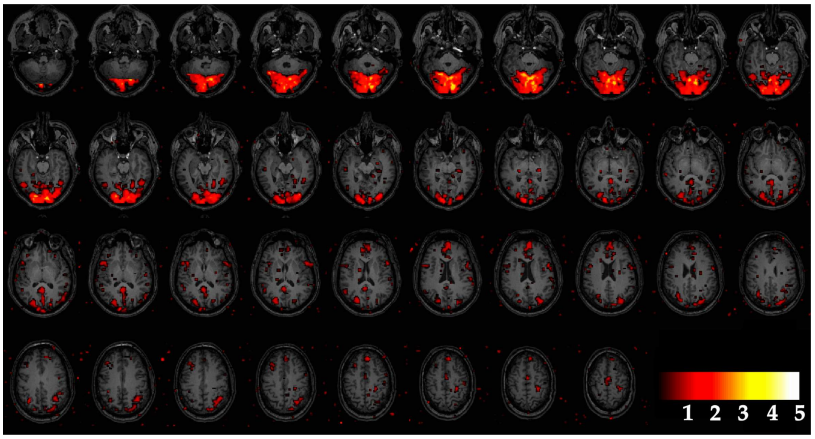


Figure 5.2: **Visual activation.** The figures show axial slices of the anatomical volume with the paradigm model order overlayed (number of harmonics of the periodic stimulus cycle). Note that no thresholding is applied to the images - it simply shows how many orders of the stimulus cycle was detected to be present in the signal. Activation is mainly restricted to the primary visual areas as is to be expected. However, as the detection scheme is extremely sensitive effects in other areas are also detected. Using this method where we simply select the most probable model we have implicitly assumed that the cost of any misclassification (model to model or model to null hypothesis) is the same. If false positives are a primary concern it may be useful to assign a higher cost to misclassification of the null hypothesis.

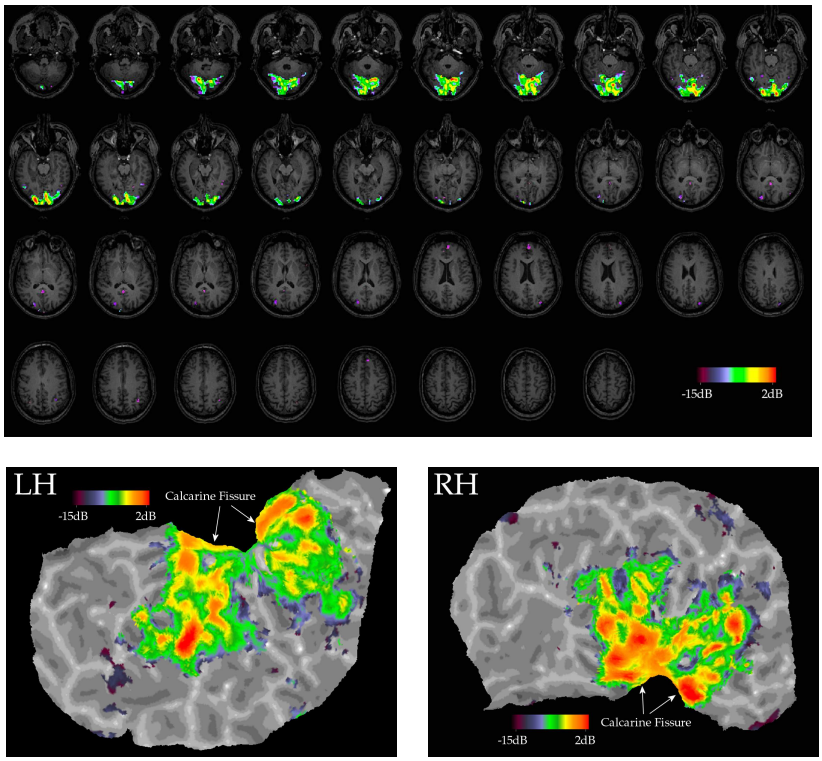


Figure 5.3: **Posterior** SNR**.** The upper panel shows the anatomical volume with the detected SNR ratio overlayed as indicated by the colour bar. The two bottom panels show flattened views of the occipital lobe with the detected posterior SNR overlayed in the colours indicated by the colour bars. LH denotes the left hemisphere whereas RH indicates the right hemisphere, in order to flatten the surfaces cuts were made along the Calcarine fissures as indicated by the arrows.

5.2 Simple Algorithms Made Complex

Despite the title which may appear intimidating this section concerns a very simple method for combining optimisation in the time and frequency domain. Frequency representation of time series data accommodate efficient optimisation for models with shifts, convolutions and smoothness due to the following basic properties

- Shift in the time/spatial domain corresponds to multiplication in the frequency domain.
- Convolution in time/spatial domain also becomes multiplication in the frequency domain.
- Smoothness of signals in the time/spatial domain correspond to low frequency representations in the frequency domain.

Further, the existence of efficient Fast Fourier Transform (FFT) algorithms allow transformation between the time/spatial and frequency domain to be performed at limited computational cost $\mathcal{O}(n \log n)$. Unfortunately, models are often constrained in the time domain in a way that is not transparent in a frequency representation which hampers optimisation in the frequency domain. In particular, component identification can be improved by imposing constrains such as sparseness, non-negativity or smoothness. While shifts and convolution are efficiently implemented in the frequency domain constraints in the form of non-negativity, sparseness and smoothness are typically defined in the time domain without an explicit representation in the frequency domain. In Mørup *et al.* (2007b) we applied the time-frequency gradient method (TFGM) which we later formalised in Madsen and Hansen (2008).

We consider objective functions, C, of the form

$$C = \sum_t f_t(x_t) + \frac{1}{N} \sum_f g_f(\widetilde{x}_f), \tag{5.12}$$

where f_t and g_f are real valued functions of the real and complex variables x_t and $\widetilde{x}_f$ such that $\widetilde{\boldsymbol{x}} = \mathcal{F}(\boldsymbol{x})$. Thus, we require the objective functions to be separable in either the time/spatial or frequency domain. The gradient with

respect to x_t and $\widetilde{x}_f$ of objective functions satisfying (5.12) can be written as

$$\begin{aligned}
\frac{\partial C}{\partial x_t} &= f_t'(x_t) + \frac{1}{N}\sum_f g_f'(\widetilde{x}_f)e^{\imath 2\pi f t} \\
&= f_t'(x_t) + \mathcal{F}^{-1}(\boldsymbol{g}')_t, \\
\frac{\partial C}{\partial \widetilde{x}_f^*} &= \sum_t f'(x_t)e^{-\imath 2\pi f t} + \tfrac{1}{N} g_f'(\widetilde{x}_f) \\
&= \mathcal{F}(\boldsymbol{f}')_f + \tfrac{1}{N} g_f'(\widetilde{x}_f).
\end{aligned}$$

Thus, the gradient of the objectives can be converted arbitrarily between the time and frequency domain. The crux of this property follows from the separability into sums over time/space or frequency instances (t or f). Further, least squares objective in a form satisfying (5.12) are equivalent in the time and frequency domain due to Parseval's identity, i.e.

$$\sum_n \|\boldsymbol{x}_n\|_F^2 = \frac{1}{N}\sum_f \|\boldsymbol{x}_f\|_F^2.$$

We note that a variable which is updated in the frequency domain has to remain real when applying the inverse DFT. For this requirement to be fulfilled, e.g. $\mathcal{F}^{-1}(\boldsymbol{g})$ real valued, the following relation has to hold in the frequency domain

$$g_{N-f+1} = g_f^*, \tag{5.13}$$

where * denotes complex conjugate. The constraint is enforced by only considering the $\lfloor N/2 \rfloor + 1$ frequencies, i.e. frequencies up to the Nyquist frequency, while setting the functions of the remaining frequencies according to (5.13).

In Madsen and Hansen (2008) we applied the TFGM to derive simple algorithms for the following FA type decompostions:

- **Shifted Non-negative Matrix Factorisation**
  The data and model parameters are constrained non-negative in the time domain while temporal shifts are represented efficiently through the frequency domain representation.
- **Convolutive Sparse Coding**
  While convolution is efficiently implemented through multiplication in the frequency domain the filter length in the time domain constrains regions of $\boldsymbol{A}^\tau$ to zero. Furthermore, the sparseness imposed on $\boldsymbol{S}$ resides in the time domain.
- **Sparse and Smooth Matrix Factorisation**
  While smoothness constraints can efficiently be implemented in the frequency domain, sparseness constraints reside in the time domain.

The following sections will briefly describe the derivation of these models.

5.2.1 Shifted Non-negative Matrix Factorisation

A popular approach for enforcing non-negativity is the use of the multiplicative updates algorithm introduced in Lee and Seung (1999, 2000). See appendices A.2.2 and A.2.3 for further details.

The shifted non-negative matrix factorisation (SHIFTNMF) model proposed in Mørup *et al.* (2007b) is given by

$$y_{m,n} \approx \sum_d a_{m,d} s_{d,n-\tau_{m,d}},$$

where $\boldsymbol{Y}$,$\boldsymbol{A}$ and $\boldsymbol{S}$ are non-negative. While shifts correspond to simple multiplication of a complex phase, the non-negativity constraint is not transparent in the frequency domain. Thus, a method combining the apparent representation of non-negativity in the time domain with the efficient implementation of shifts in the frequency domain is desired. The least squares objective can be written as

$$C_{LS}(A,S) = \frac{1}{2N} \sum_f \|\widetilde{\boldsymbol{y}}_f - \widetilde{\boldsymbol{A}}^{(f)} \widetilde{\boldsymbol{s}}_f\|_F^2.$$

Thus, in the frequency domain the objective becomes separable over frequencies, however, the non-negativity constraint resides in the time domain. The model can be estimated alternatingly solving for $\boldsymbol{A}$, $\boldsymbol{S}$ and $\boldsymbol{\tau}$ as described in Mørup *et al.* (2007b). Here, the TFGM is applied when updating the variable $\boldsymbol{S}$. The gradient of the least squares cost function in the frequency domain is

$$\widetilde{\boldsymbol{g}}_f = \frac{\partial C_{LS}}{\partial \widetilde{\boldsymbol{S}}_f} = -\frac{1}{N} \widetilde{\boldsymbol{A}}^{(f)^H} (\widetilde{\boldsymbol{x}}_f - \widetilde{\boldsymbol{A}}^{(f)} \widetilde{\boldsymbol{s}}_f).$$

By applying the inverse DFT on the gradient in the frequency domain the corresponding gradient in the time domain is obtained. Splitting the gradient in the frequency domain into what constitutes the positive and negative part of the corresponding gradient in the time-domain gives

$$\begin{aligned} \widetilde{\boldsymbol{g}}_f^+ &= \frac{1}{N} \widetilde{\boldsymbol{A}}^{(f)^H} \widetilde{\boldsymbol{A}}^{(f)} \widetilde{\boldsymbol{s}}_f, \\ \widetilde{\boldsymbol{g}}_f^- &= \frac{1}{N} \widetilde{\boldsymbol{A}}^{(f)^H} \widetilde{\boldsymbol{y}}_f. \end{aligned}$$

Consequently, by taking the inverse DFT of $\widetilde{\boldsymbol{G}}^+$ and $\widetilde{\boldsymbol{G}}^-$ the corresponding positive and negative part of the gradient in the time-domain are found. As a result,

$\boldsymbol{S}$ can be updated using multiplicative updates in the time domain, hence, enforcing non-negativity through the update

$$s_{d,n} = s_{d,n} \left(\frac{g^{-}_{d,n}}{g^{+}_{d,n}} \right)^{\alpha}.$$

In figure 5.4 we demonstrate the usefulness of the SHIFTNMF over regular instantaneous NMF when shifts are present in the data.

5.2.2 Convolutive Sparse Coding

Traditionally, convolutive BSS models have been estimated either in the time domain by Toeplitz structured matrices or in a windowed frequency domain representation Syskind *et al.* (2007). We will presently solve the convolutive BSS problem in the time domain using the TFGM. The Convolutive Sparse Coding model is given by

$$y_{m,n} \approx \sum_{d,\tau} a^{\tau}_{m,d} s_{d,n-\tau}.$$

Where $\boldsymbol{S}$ is sparse. The model is separable in the frequency domain and can be optimised using the following objective of the form given in (5.12)

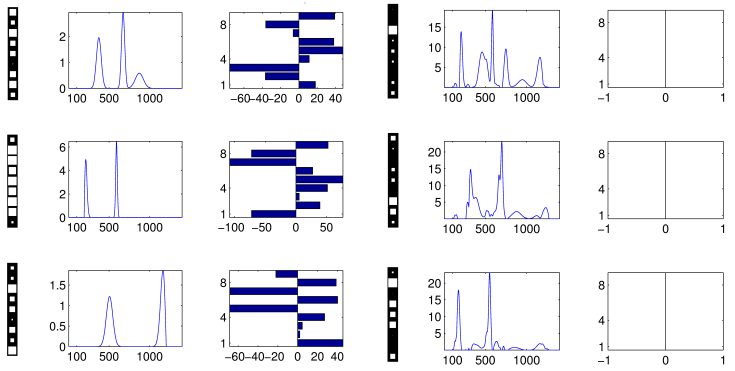
$$C = \frac{1}{2N} \sum_{f} \|\widetilde{\boldsymbol{y}}_f - \widetilde{\boldsymbol{A}}_f \widetilde{\boldsymbol{s}}_f\|^2_F - \lambda \sum_{n} \log(sp(\boldsymbol{s}_n)).$$

Where the first term is the reconstruction error and second term the sparsity penalty imposed with strength λ given by the sparse prior distribution sp. We will consider the Laplace prior given by $sp(\boldsymbol{s}_n) = e^{-|\boldsymbol{s}_n|}$ forming a l_1-norm regularisation penalty. The sparsity in the time domain as well as regions where the filter $\boldsymbol{A}^{\tau}$ is zero is not transparent in a frequency domain representation. However, the convolutive model is efficiently estimated in a frequency domain representation. Thus, again the TFGM admits the benefits of the representations in the two domains. The gradient of the least squares error in the frequency domain is given by

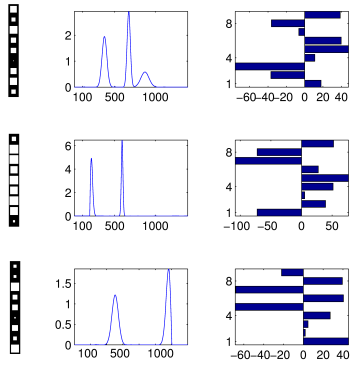
$$\begin{aligned} \nabla^{LS}_{\widetilde{a}_{m,d,f}} &= -\frac{1}{N}(\widetilde{y}_{m,f} - \sum_{d} \widetilde{a}^{*}_{m,d,f} \widetilde{s}_{d,f}) \widetilde{s}^{*}_{d,f} \\ \nabla^{LS}_{\widetilde{s}_{d,f}} &= -\frac{1}{N} \sum_{m} \widetilde{y}_{m,d,f} (\widetilde{y}_{m,f} - \sum_{d} \widetilde{a}^{*}_{m,d,f} \widetilde{s}_{d,f}) \end{aligned}$$

Thus, the gradient in the time domain is given by

$$\begin{aligned} \nabla_{\boldsymbol{A}^{\tau}} &= \mathcal{F}^{-1}(\nabla^{LS}_{\widetilde{\boldsymbol{A}}})_{\tau} \\ \nabla_{\boldsymbol{S}} &= \mathcal{F}^{-1}(\nabla^{LS}_{\widetilde{\boldsymbol{S}}}) + \lambda \operatorname{sign}(\boldsymbol{S}) \end{aligned}$$



(a) True components (b) Estimated components NMF



(c) Estimated components SHIFTNMF

Figure 5.4: **Upper left panel:** The true factors forming the synthetic data ($\boldsymbol{Y} \in \Re^{9\times 1400}$). To the left, the strength of the mixing $\boldsymbol{A}$ of each source is indicated in gray color scale. In the middle, the three sources are shown and to the right is given the time delays of each source to each channel. **Upper right panel:** Results obtained by conventional instantaneous NMF for the generated synthetic data. Clearly, the model cannot account for the shifts in the data hence the sources estimated are incorrectly estimated. Notice, only 68 % of the variance of the data can be accounted for. **Bottom panel:** The estimated factors obtained by a SHIFTNMF analysis. Clearly, the model with shifts has correctly recovered the components of the synthetic data hence accounts for all the variance in the data.

Hence, by computing the gradient in the time domain it becomes transparent how $\boldsymbol{A}^\tau$ can be estimated such that only active regions of the filter $\boldsymbol{A}^\tau$ are updated. Notice, how the complexity of this approach does not increase when increasing the filter length. Furthermore, the update in the time domain of $\boldsymbol{S}$ enables the combination of sparseness constraint in the time domain with efficient representation in the frequency domain. $\boldsymbol{A}^\tau$ and $\boldsymbol{S}$ are updated using linesearch, i.e. by $\boldsymbol{A}^\tau \leftarrow \boldsymbol{A}^\tau - \mu_A \nabla_{\boldsymbol{A}^\tau}$ and $\boldsymbol{S} \leftarrow \boldsymbol{S} - \mu_S \nabla_{\boldsymbol{S}}$.

In Madsen and Hansen (2008) we demonstrated this algorithm on EEG data.

5.2.3 Sparse and Smooth Matrix Factorisation

Smoothness is often imposed by penalising changes over time in the signal based on penalty terms such as local estimates of the temporal derivative, i.e. $\|s_t - s_{t-1}\|_F^2$ or curvature $\|s_{t-1} + s_{t+1} - 2s_t\|_F^2$ as described in Bro (1998); Hastie and Tibshirani (1990). In the frequency domain this can be expressed as

$$\|\widetilde{w}_f \boldsymbol{s}_f\|_F^2 \tag{5.14}$$

where $\widetilde{w}_f = 1 - e^{-i2\pi f/N}$ and $\widetilde{w}_f = 2 - e^{-\imath 2\pi f/N} - e^{i2\pi f/N}$ for the derivatives above. Smoothness constraints in the time domain correspond to reduced high frequency content. Hence, smoothness can be imposed by penalising high frequency regions of the components, i.e. by considering an objective of the form given in (5.12)

$$C = \sum_n \frac{1}{2}\|\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n\|_F^2 + \frac{\lambda_2}{2N}\sum_f \|\widetilde{w}_f \widetilde{\boldsymbol{s}}_{\boldsymbol{f}}\|_F^2.$$

From the objective above it can be seen that smoothness does not improve the identifiability of the model since multiplying the sources $\boldsymbol{S}$ by the orthogonal matrix $\boldsymbol{Q}$ result in a representation that is equally smooth, i.e. $\|\widetilde{\boldsymbol{s}}_{\boldsymbol{f}}\|_F = \|\boldsymbol{Q}\widetilde{\boldsymbol{s}}_{\boldsymbol{f}}\|_F$. Thus, additional constraints are necessary in order to obtain an unambiguous representation. We will here improve the identifiability of the model by imposing sparseness on $\boldsymbol{S}$. Again, sparsity is not transparent in a frequency representation. However, the sparsity and smoothness constraints can again be combined using the proposed TFGM. Consider the following sparse and smooth matrix factorisation

$$C = \frac{1}{2}\sum_n (\|\boldsymbol{y}_n - \boldsymbol{A}\boldsymbol{s}_n\|_F^2 + \lambda_1 \|\boldsymbol{s}_n\|_1) + \frac{\lambda_2}{2N}\sum_f \|\widetilde{w}_f \widetilde{\boldsymbol{s}}_{\boldsymbol{f}}\|_F^2,$$

where $\widetilde{w}_f$ weights frequencies according to the smoothness desired. Clearly, the objective has the form given in (5.12). Thus, the gradient of the above objective

is given by

$$\begin{aligned}\nabla_{\boldsymbol{A}} &= -(\boldsymbol{AS}-\boldsymbol{Y})\boldsymbol{S}^{\top}\\ \nabla_{\boldsymbol{S}} &= -\boldsymbol{A}^{\top}(\boldsymbol{AS}-\boldsymbol{X})+\lambda_1\,\mathrm{sign}(\boldsymbol{S})+\lambda_2\mathcal{F}^{-1}(\widetilde{\boldsymbol{S}}').\end{aligned}$$

where, $\boldsymbol{s}'_f|\widetilde{w}_f\widetilde{w}_f^*\boldsymbol{s}_f$. Again $\boldsymbol{A}$ and $\boldsymbol{S}$ are updated using linesearch, i.e. by $\boldsymbol{A}\leftarrow \boldsymbol{A}-\mu_A\nabla_{\boldsymbol{A}}$ and $\boldsymbol{A}\leftarrow \boldsymbol{S}-\mu_S\nabla_{\boldsymbol{S}}$.

In figure 5.5 we demonstrate how smoothness and sparseness imposed on simulated data improves identifiability and reduces noise in the decomposition.

Similarly spatial smoothness of components can be imposed by penalising high frequency spatial factors. For fMRI data which is usually recorded in three dimensions this requires that the DFT is performed on arrays that respect the original spatial structure of the data (e.g. a three-way array). It is typically not appropriate to enforce smoothness on edges of the brain (borders of the brain mask). Unfortunately, it is not trivial to exclude regions or to use windowing functions when applying the TFGM and it is often easier to calculate the cost function and gradient in the spatial domain, where it is easy to exclude edges/regions, instead of in the frequency domain. A cost function often used to measure spatial smoothness is the Laplacian based on curvature (here written with respect to the Cartesian coordinates l,m,n of the function $f(l,m,n)$).

$$\nabla^2 f=\frac{\partial^2 f}{\partial l^2}+\frac{\partial^2 f}{\partial m^2}+\frac{\partial^2 f}{\partial n^2}$$

We can approximate the Laplacian locally based on finite differences of the six immediate neighbouring points (6 neighbourhood) which corresponds to the curvature penalty.

$$\begin{aligned}L(l,m,n) &= f(l+1,m,n)-f(l-1,m,n)+f(l,m+1,n)\\ &\quad -f(l,m-1,n)+f(l,m,n+1)-f(l,m,n-1)\\ &\quad -6f(l,m,n)\end{aligned}$$

This can straightforwardly be generalised to larger neighbourhoods (such as the 26 neighbourhood that is used in the generation of figure 5.6). When the cost function is written in the spatial domain it is relatively simple to exclude effects from edges in calculation of the cost function and gradient. In figure 5.6 we show how spatial smoothness regularisation based on the l_2-norm can suppress noise in a SC decomposition. Spatial regularisation based on the l_2-norm does not render the decomposition unique, an attractive alternative is l_1-norm regularisation. However, if gradient methods are used naively to optimise parameters severe convergence problems can occur due to elements oscillating around zero (these elements are supposed to be zero) causing step sizes to become arbitrarily

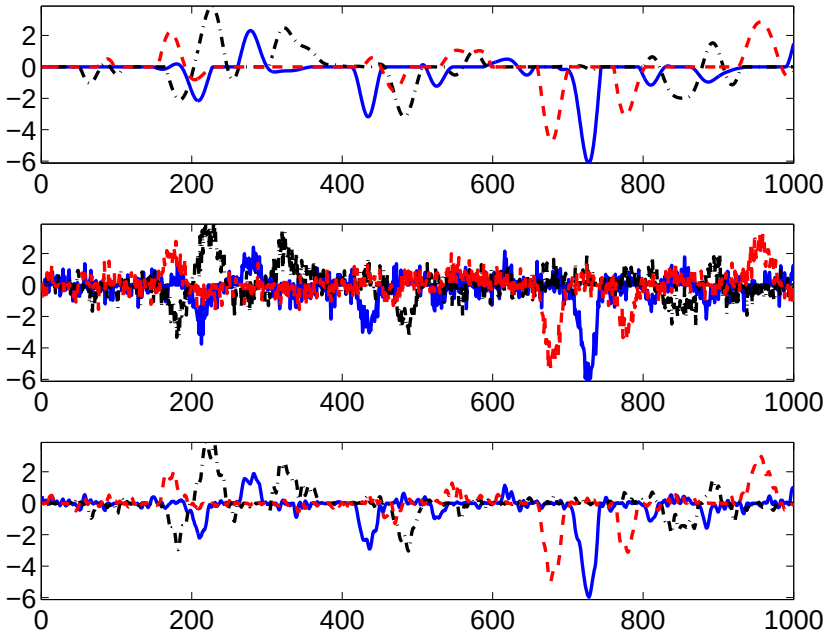


Figure 5.5: **Sparse and smooth matrix factorisation.** The top graph shows the three true underlying sources used to simulate data, all three components are both sparse and smooth. These sources was mixed such the the input consisted of nine realisations of the signal ($\mathbf{Y} \in \Re^{9\times 1000}$) with additive Normal distributed noise (SNR of 0 dB). The middle graph shows the sources identified by the proposed smooth and sparse matrix factorisation algorithm imposing ($\lambda_1 = 0.25$, $\lambda_2 = 0$, i.e., without any smoothness regularisation). Finally, the bottom graph shows how introducing smoothness regularisation suppresses noise in the solution ($\lambda_1 = 0.25$, $\lambda_2 = 20$ for frequencies larger than 0.14 samples^{-1}). While smoothness suppresses noise in the sources it will also introduce a bias in this particular case reducing the explained variation from 64% to 53%.

small. One possible solution is using an IRLS scheme (see for example Blumensath and Davies (2004)) for estimation. However, this iterative scheme seem to suffer from very slow convergence for this particular application making it impractical to use it for large-scale data sets. An adapted version of the robust gradient method described in A.2.1 could be employed in the spatial frequency domain to circumvent oscillation problems in gradient based optimisation, providing a more efficient way to optimise parameters (however, in this case it is not trivial to completely circumvent the edge effects described earlier).

5.3 Shift Invariant Multi-way Decomposition

In Mørup *et al.* (2008b) we present an algorithm allowing multi-way decomposition with explicit modelling of delays over one mode. The model is motivated by the fact that time delays are known to occur naturally in neuroimaging data.

As proved by Kruskal (1977), the regular instantaneous CP model is unique under mild conditions. Conditions that, in the presence of noise in the data, are practically always satisfied. Consequently, modeling repeated trials by CP in theory not only improves the component identification but also resolves the ambiguities encountered when modeling the averaged data by (bilinear) factor analysis. Notice that the application of CP to EEG was already suggested in the original paper on CP (Harshman, 1970) and was later reinvented in Möcks (1988) under the name *topographic component analysis*. In Andersen and Rayens (2004) it was further demonstrated how the CP model is useful in the analysis of neuroimaging data such as fMRI (Andersen and Rayens, 2004). Additional applications of multilinear (also called multiway) modeling in EEG and fMRI include (Möcks, 1988; Field and Graupe, 1991; Wang *et al.*, 2000; Beckmann and Smith, 2005; Miwakeichi *et al.*, 2004; Mørup *et al.*, 2006; De Vos *et al.*, 2007; Acar *et al.*, 2007).

Time shifts occur naturally in fMRI data. For instance, these could be due to hemodynamic delay (Buxton *et al.*, 1998) or they could arise in stimuli studies (Sereno *et al.*, 1995), where delays play a particularly important role. Extending the CP model to incorporate delays form the shifted CP model (shifted over third mode), denoted as SCP model,

$$x_{i,j,k} = \sum_{d=1}^{D} a_{i,d} b_{j-\tau_{k,d},d} c_{k,d} + e_{i,j,k}. \tag{5.15}$$

Here, each time profile $\boldsymbol{b}_d$ is shifted according to the vector $\tau_{k,d}$ that represents time-samples dependent on the k index of the third mode. Hence, the shifts will be along the j index, see also Figure 5.7. Data generated from the SCP model

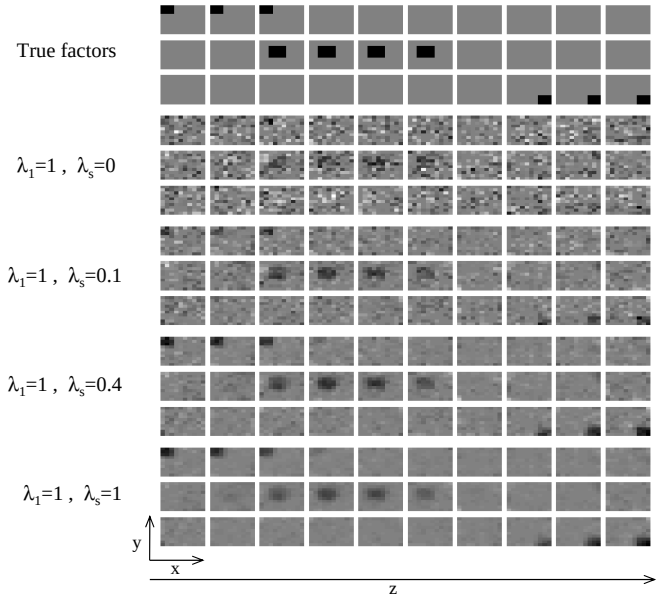


Figure 5.6: **Imposing spatial smoothness and sparseness.** The figure illustrates how smoothness regularisation (here in three dimensions denoted by x, y and z) when used with sparsity imposed through L_1-norm regularisation may improve the identification of components. In the figure above λ_1 indicates the strength of the sparsity regularisation whereas λ_s refer to the strength of the smoothness regularisation. The (spatial) dimensionality of the 3 factors are $10 \times 10 \times 10$ for visualisation purposes the third dimension has been unfolded (along the x-axis). Each row of images therefore corresponds to an unfolded factor. From top to bottom the true factors along with several different attempts to find the factors are displayed each consisting of 3 factors. The input data was simulated by a mixing matrix of size 50×3 (Gaussian with scalar covariance) which mixed the factors into a 50×1000 data matrix with additive Gaussian noise (also scalar covariance). A low SNR of approximately -18.5 dB was chosen at which the sparseness prior was not sufficient to produce satisfactory factors. Notice how increasing smoothness suppresses noise making it possible to identify the true factors but makes it more difficult to identify the non-smooth edges in the images (here the smoothness penalty of $\lambda_s = 0.4$ seems to produce the best result).

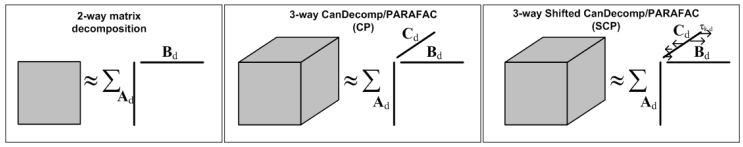


Figure 5.7: The CP model can be considered a straightforward generalisation of 2-way (matrix) decomposition (left panel) to arrays of more than two modalities (middle panel). Thus, the data is described by an outer product of factors pertaining to each of the modalities. The SCP model allow shifts to occur over the second mode such that for each index of the third mode the component of the second mode is shifted a given amount.

is no longer multilinear and therefore the CP model is no longer a valid model for the data. When data violates multi-linearity, 'CP-degenerate' solutions are known to occur. Roughly speaking, this refers to solutions in which some component loadings are highly correlated in all modes and the elements of these components become arbitrarily large (Stegeman, 2007). CP-degeneracy makes the estimation unstable, the algorithm slow to converge (or even diverge), and the result difficult to interpret — largely because the model is plaqued by strong between-component cancellations (Harshman and Lundy, 1984). To avoid CP-degeneracy in the CP model, artificial restrictions in the form of orthogonality (Harshman and Lundy, 1984; Field and Graupe, 1991) or independence (Beckmann and Smith, 2005) have been imposed; alternatively, the signal is analysed via purely additive models based on analysis of amplitudes in a spectral representation (Miwakeichi *et al.*, 2004; Mørup *et al.*, 2006). Rather than restricting the CP model, we propose a *pseudo-multilinear* model using the unambiguous CP model combined with a time-shift accounting for explicit delays.

Modeling of delays is further motivated by a number of papers that explain how degenerate solutions might be caused by component delays (Field and Graupe, 1991; Andersen and Rayens, 2004; Harshman *et al.*, 2003a; Hong and Harshman, 2003). Indeed, if shifts are causing CP-degeneracy, then it is more natural to extend the CP model to accomodate shifts rather than resorting to orthogonality or independence constraints that may not be physiologically justified. Furthermore, a decomposition into profiles resembling pairs of functions and their derivatives, e.g. pairs of cosine and sine functions in (Field and Graupe, 1991), provides strong evidence that neuroimaging data should be decomposed by a model accounting for shifts rather than models based on instantaneous mixing.

Methods

Factor analysis with shifts have been treated in numerous papers (Bell and Sejnowski, 1995; Harshman *et al.*, 2003a,b; Truccolo *et al.*, 2003; Mørup *et al.*, 2007a,b). Shifts based on the CP model has previously been treated in (Hong and Harshman, 2003; Knuth *et al.*, 2006). Unfortunately, the algorithms devised are prohibitively slow for large scale problems such as EEG and fMRI and do not allow for non-integer shifts. Presently, we derive an efficient algorithm for SCP with the following benefits

- Closed form solutions are obtained for all modes while keeping the remaining modes fixed.
- Integer shifts are estimated by cross-correlation rather than the exhaustive searches used in (Hong and Harshman, 2003).
- Non-integer shifts can be found by iterative methods in the frequency domain.

Non-integer shifts are in particular important for fMRI data due to low temporal resolution.

Estimating the SCP model

In (Hong and Harshman, 2003) the SCP model was proposed and an algorithm devised based on exhaustive integer searches over all possible shifts. This is, however, very expensive making the estimation infeasible when including many shifts. Thus, we here propose to solve the model in the frequency domain rather than the time-domain. The attractive property being that each integer delay $\tau_{k,d}$ has a closed form solution while keeping the remaining delays fixed given by calculating cross-correlations which is inexpensive in the frequency domain. Furthermore, in a frequency representation non-integer delays can be estimated using gradient based searches. Finally, in a frequency representation $\boldsymbol{B}$ has a closed form solution.

In the frequency domain the SCP model is given by

$$x_{i,f,k} = \sum_{d=1}^{D} a_{i,d}\tilde{b}_{f,d}c_{k,d}\exp[-\imath 2\pi\frac{f-1}{J}\tau_{k,d}] + \tilde{e}_{i,f,k}.$$

Thus, the sources $\boldsymbol{b}_d$ are assumed to be periodic such that shifts $\tau_{k,d}$ correspond to the complex multiplication of $\tilde{\boldsymbol{b}}_d$ with the factor $\exp[-\imath\pi\frac{f-1}{J}\tau_{k,d}]$. Thus, we

assume that the data can be arranged such that each source time course in each epoch is periodic, if this is not the case the periodicity can be enforced by introducing a temporal windowing function. Notice, due to Parseval's identity there is a one-to-one correspondence between the least squares error in the time and frequency domain such that the least squares minimisation can be performed arbitrarily between the two domains

$$\sum\nolimits_{i,j,k} \|e_{i,j,k}\|^2 = \frac{1}{J} \sum\nolimits_{i,f,k} \|\tilde{e}_{i,f,k}\|^2.$$

5.3.0.1 $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ updates

Let $\tilde{b}^{(k)}_{f,d} = \tilde{b}_{f,d} \bullet \exp[-\imath 2\pi \frac{f-1}{J} \tau_{k,d}]$, i.e. $\tilde{\boldsymbol{B}}$ componentwise shifted according to the delays to the k^{th} channel. Let further $z_{j+k(J-1),d} = c_{k,d} b^{(k)}_{j,d}$, i.e. the Khatri-Rao product between $\boldsymbol{C}$ and the shifted version of $\boldsymbol{B}$.

Using n-mode matricising and the Khatri-Rao product we can state the estimation of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ by ordinary factor analysis

$$\begin{aligned}
\boldsymbol{X}_{(1)} &= \boldsymbol{A}\boldsymbol{Z}^{\mathsf{T}} + \boldsymbol{E}_{(1)} \quad \text{via} \quad \boldsymbol{A} \leftarrow \mathbf{X}_{(1)} \boldsymbol{Z}^{\mathsf{T}\dagger} \\
\tilde{\boldsymbol{x}}_{(2)_{f,:}} &= \tilde{\boldsymbol{b}}_{f,:} (\tilde{\boldsymbol{C}}^{(f)} \odot \boldsymbol{A})^{\mathsf{T}} + \tilde{\boldsymbol{e}}_{(2)_{f,:}} \quad \text{via} \quad \tilde{\boldsymbol{b}}_{f,:} \leftarrow \tilde{\boldsymbol{x}}_{(2)_{f,:}} (\tilde{\boldsymbol{C}}^{(f)} \odot \boldsymbol{A})^{\mathsf{T}\dagger} \\
\boldsymbol{x}_{(3)_{k,:}} &= \boldsymbol{c}_{k,:} (\boldsymbol{B}^{(k)} \odot \boldsymbol{A})^{\mathsf{T}} + \boldsymbol{e}_{(3)_{k,:}} \quad \text{via} \quad \boldsymbol{c}_{k,:} \leftarrow \mathbf{X}_{(3)_{k,:}} (\boldsymbol{B}^{(k)} \odot \mathbf{A})^{\mathsf{T}\dagger}
\end{aligned}$$

Notice, where as $\boldsymbol{A}$ and $\boldsymbol{C}$ are updated in the real domain $\boldsymbol{B}$ is updated in the complex domain.

5.3.0.2 $\boldsymbol{\tau}$ update

Let

$$\boldsymbol{r}_{(3)^{d'}_{k,:}} = \boldsymbol{x}_{(3)_{k,:}} - \sum\nolimits_{d \neq d'} c_{k,d} (\boldsymbol{b}^{(k)}_d \odot \boldsymbol{a}_d)^{\mathsf{T}},$$

i.e. $\boldsymbol{r}^{d'}_{(3)_{k,:}}$ is the remaining signal at the k^{th} row when projecting all but the d'^{th} source out of $\boldsymbol{X}_{(3)}$. Notice, with this notation the least squares error can be rewritten as

$$\begin{aligned}
&\sum\nolimits_k \|\boldsymbol{x}_{(3)_{k,:}} - \sum\nolimits_d^D c_{k,d} (\boldsymbol{b}^{(k)}_d \odot \boldsymbol{a}_d)^{\mathsf{T}})\|^2 \\
=&\sum\nolimits_k \|\boldsymbol{r}^{d'}_{(3)_{k,:}} - c_{k,d'} (\boldsymbol{b}^{(k)}_{d'} \odot \boldsymbol{a}_{d'})^{\mathsf{T}}\|^2 \\
=&\|\boldsymbol{r}^{d'}_{(3)_{k,:}}\|^2 - c_{k,d'} \sum\nolimits_j b_{j-\tau_{k,d'},d'} \sum\nolimits_i r^{d'}_{i,j,k} a_{i,d'} + \|c_{k,d'} (\boldsymbol{b}^{(k)}_{d'} \odot \boldsymbol{a}_{d'})^{\mathsf{T}}\|^2.
\end{aligned}$$

The first and third term is independent of $\tau_{k,d'}$. Thus, the least square error is minimised when the second term is maximised. Let,

$$s_{k,d'}(j) = \sum\nolimits_i r_{i,j,k}^{d'} a_{i,d'}$$
$$\tilde{v}_{k,d'}(f) = \tilde{s}^*_{k,d'}(f)\tilde{b}_{f,d'}.$$

$\tau_{k,d'}$ can now be estimated as

$$\hat{\tau}_{k,d'} = \underset{t}{\operatorname{argmax}}|v_{k,d'}(t)|$$
$$\tau_{k,d'} = \hat{\tau}_{k,d'} - (J+1).$$

I.e. as the delay corresponding to maximum absolute cross-correlation between $s_{k,d'}(j)$ – the time profile of the residual for the d' component and $b_{d'}$-the component time profile. The value of $c_{k,d'}$ corresponding to this delay is given by

$$c_{k,d'} = \frac{z_{k,d'}(\hat{\tau}_{k,d'})}{\boldsymbol{b}_{d'}^\top \boldsymbol{b}_{d'}}.$$

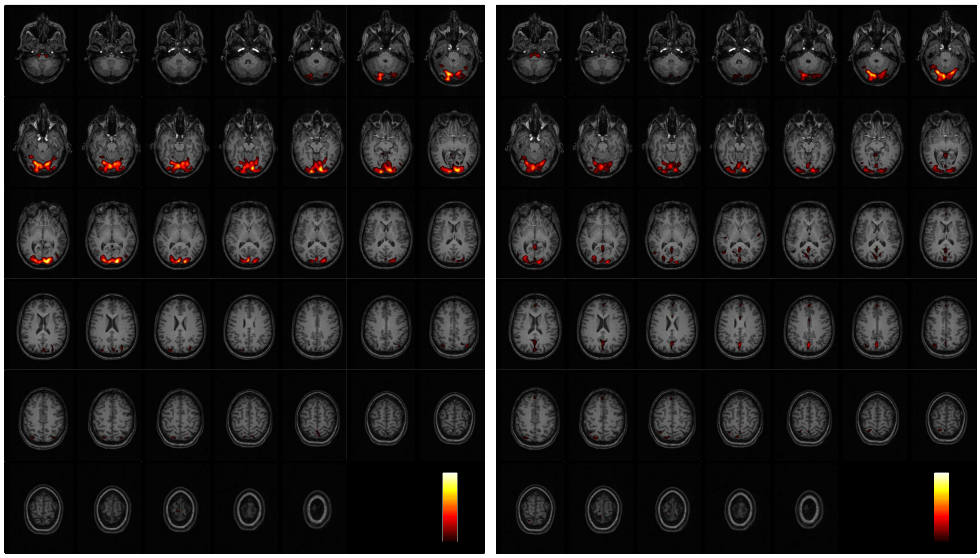
If $\mathbf{C}$ is constrained positive only positive values of $z_{k,d}(t)$ are considered. The above procedure can only estimate integer delays. However, by minimising the least squares error in the complex domain with respect to $\boldsymbol{\tau}$ a gradient and Hessian can be calculated such that non-integer delays can be estimated for instance by the Newton-Raphson procedure.

In Mørup *et al.* (2008b) we demonstrate the model on simulated and real EEG data as well as fMRI data. In figure 5.8 the results using the algorithm on a standard fMRI retinotopic mapping paradigm are presented. Figure 5.8 gives the estimated spatial activity ($\boldsymbol{A}$) found by SCP for a standard ring and wedge paradigm. In addition figure 5.9 gives the estimated delays.

The delays observed correspond well to delays obtained from a traditional voxel based DFT analysis (Sereno *et al.*, 1995; Engel *et al.*, 1997; Warnking *et al.*, 2002). Where it is assumed that the time series for the delay modeling is sinusoidal and constant in strength over the epochs. Thus, the benefits of the SCP are that noisefull epochs are given less importance in the estimation of the delays while a more complex pattern of the time series improves the delay estimation. For further details on the model, results and discussion see Mørup *et al.* (2008b).

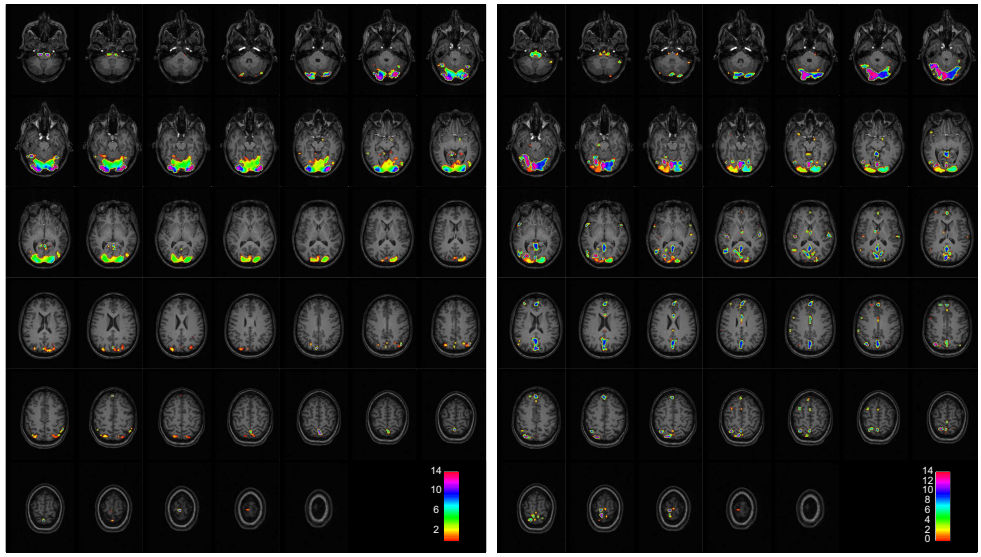
5.4 Slice-wise Modelling of fMRI Data

An often used approach for dealing with different timings for each slice in fMRI data is to interpolate the signal in time to make it appear as if all slices were



(a) Ring: Est. voxel Act. **C** (SCP) (b) Wedge: Est. voxel Act. **C** (SCP)

Figure 5.8: The component strength (**C**) over voxels overlayed on the high resolution structural scan for the one component SCP model. The map was threshold such that the 5% of the voxels with the largest voxel score **C** are shown, a standard Z-transform is not meaningful because **C** is constrained non-negative. **Top left panel:** Clearly, the most prominent activity found by SCP on the ring paradigm corresponds well with areas pertaining to visual information processing, i.e. visual cortex. **Top right panel:** Also for the wedge paradigm the most prominent activity for the one component SCP model pertains to visual cortex.



(a) Ring: Est. delays $\boldsymbol{\tau}$ (SCP)

(b) Wedge: Est. delays $\boldsymbol{\tau}$ (SCP)

Figure 5.9: The estimated temporal delays based on the one component SCP model as well as the traditional DFT analysis for the ring and wedge paradigms. **Top left panel:** Delays estimated by SCP for the ring paradigm. Clearly, the delays are symmetric across the two hemispheres.**Top right panel:** Delays estimated by SCP for the wedge paradigm. Clearly, there is a difference in the delays between right and left visual field, i.e. right and left hemisphere. **Bottom left and right panel:** Delays estimated based on the phases obtained by a voxel based DFT analysis of the ring and wedge paradigms. Similar symmetries are found as obtained by the SCP, however, the maps are not as smooth as the delay maps found by SCP thus appear somewhat more confounded by noise.

acquired simultaneously (Henson *et al.*, 1999). Such an approach enables the use of the same design matrix for all voxels but require that all parts of the signal are sufficiently sampled. While this may be true for the effects of interest it is typically not the case for cardiac signal components.

In order to investigate effects of different acquisition times for the slices we construct a modified NVR approach. Here we model each slice separately based on the individual timing. This (potentially) causes the design matrix to be different for each slice. This effects of difference in slice acquisition timing may also be present in the effects of interest as demonstrated by Van de Moortele *et al.* (1997). However, the use of a flexible model such as a harmonic expansion applied here allow us to capture the effects of the visual stimulus regardless of phase differences. Had this not been the case the paradigm regressors can also be shifted accordingly like it is done in the fMRIstat software package (Worsley *et al.*, 2002).

The fact that fMRI volumes are usually obtained interleaved and that we expect neighbooring slices (spatially) to be confounded by physiological noise in a similar way enables us to investigate whether non-stationarity of physiological noise within the acquisition of one volume is a concern. We will do so by calculating the histogram of voxels exhibiting cardiac effects over slices (spatially), this histogram is expected to be smooth if the assumption of stationarity holds. In figure 5.10 we show these histograms of cardiac effects colour coded by the cardiac phase using the first, middle and last slice as reference as well as for slice-wise modelling. The data used in the generation of these plots will be described in the next section (5.5). Each histogram shows the cardiac phases over the 40 slices acquired in each of 16 scanning sessions. It is seen that cardiac effects are most predominant at the lower slices near the circle of Willis and the medial cerebral artery. The colour coding reflects the phases of the cardiac cycle estimated by the first order of the RETROICOR cardiac regressors. The phases are widely distributed over the whole range ($-\pi$ to π) even in a single slice making it reasonable to assume that the features of the cardiac cycle may be captured in a single slice. Evidence for the interleaved acquisition order of the slices is most pronounced in sessions 7W and 8R. This may be caused by differences in the detection power depending on the time of sampling. From the histograms we clearly see that using either the first or the last slice as reference causes the effect of slice acquisition to be visible, by using the middle slice (time-wise) the effect is smaller but still present. When using slice-wise modelling the effect has almost disappeared indicating that slice-wise modelling of physiological noise is appropriate.

A toolbox for performed slice-wise modelling of physiological artefacts using the RETROICOR procedure was developed during investigation of these effects. This toolbox functions as a plug-in for SPM5 and also provides the possibility of

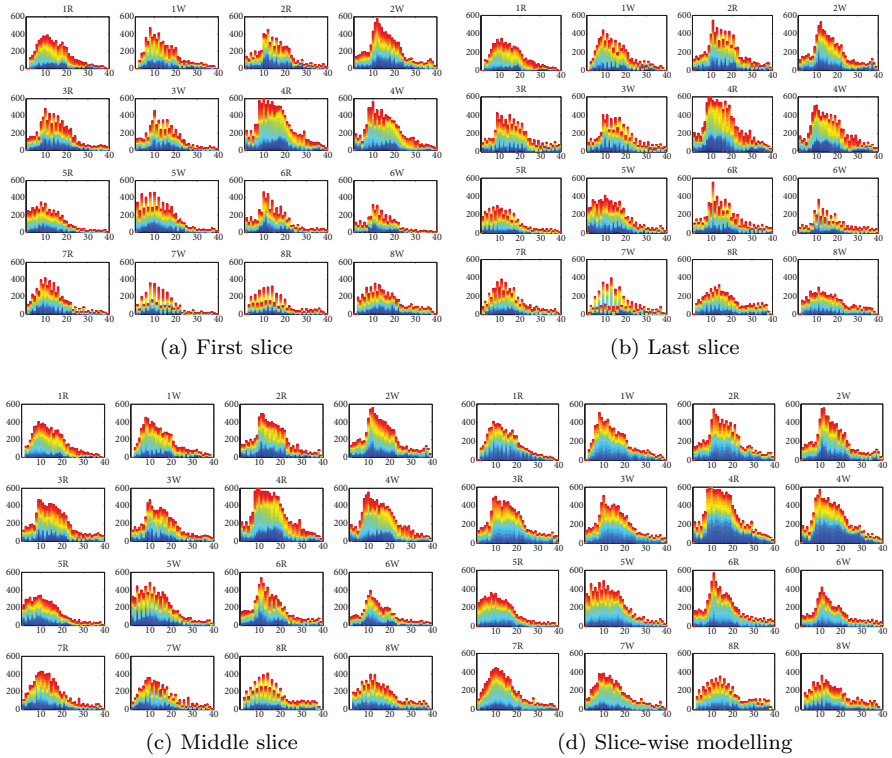
automatically calculating and adding RETROICOR regressors based on files with recordings of cardiac and respiratory cycles. In addition, the toolbox also allows slice-wise modelling of effects of interest. This is accomplished by performing model specification, estimation, AR modelling and calculation of p-values on a slice-wise basis. Note that when using this toolbox smoothing should not be performed until after modelling. A currently preliminary version of the toolbox is available at http://www.brain-fmri.com.

5.5 Unsupervised Nuisance Variable Regression

Physiological noise originating from respiration and the cardiac cycle are prominent sources of temporally correlated noise in fMRI data. These correlations violate assumptions usually made about the residual error in standard univariate general linear model analyses. Both supervised correction methods, based on recordings of the cardiac and respiratory cycles, and unsupervised correction methods, attempting to estimate the noise components from the data have been suggested. The supervised methods reduce the presence of temporal correlation in the residual noise, whereas this remains to be shown for unsupervised methods. Here we propose an unsupervised method for the correction for physiological noise contributions in fMRI data using PCA, and show that the method successfully reduces temporal correlations in the residual equivalent to supervised methods. Here, we suggest extracting information about the physiological noise from the data itself to include these effects in a standard GLM and evaluate the consequences of this approach on the temporal whiteness and normality of the residual errors.

Method

We suggest to replace the RETROICOR regressors described in section 4.5.1 by a number of nuisance regressors estimated from the data itself (hence the term unsupervised) also intended to model physiological noise contributions. This model is useful in cases where measures of the respiratory and/or cardiac cycles are not available. One way to capture nuisance effects is to use the time series of a voxel known to contain only nuisance effects directly, this approach has been demonstrated in several cases. In a study by (Petersen *et al.*, 1998) a seed region within the sinus sagittalis was used to fit a non-linear state space model and obtain phase estimates that where subsequently used to regress out related effects in a GLM. Similarly Lund and Hanson (1999) observed that the residual variance could be used to automatically identify regions related to



(a) First slice (b) Last slice

(c) Middle slice (d) Slice-wise modelling

Figure 5.10: **Cardiac phases over slices.** The figure shows histograms over slices for each of 16 scanning sessions. The four panels differ by the choice of reference slice for the RETROICOR regressors. It is seen that effects of slice acquisition are clearly visible for the three models with a single reference slice, however, for slice-wise modelling the effects appears significantly reduced. Indicating that taking slice acquisition time into account when modelling cardiac effects is appropriate.

physiological noise and include these time series in the analysis. In a study of functional connectivity by Fox *et al.* (2006) the time series of a region within a ventricle was used to remove nuisance effects before the analysis. In this study we expand this method to use time series from different regions either based on regions extracted from the high resolution anatomical scan or a parametric map of the residual variance from a model excluding these regressors. The choice of the residual variance as an identifier for nuisance regions is based on the observation that parametric maps of the residual variance closely resembles a coarse angiogram, we will denote this procedure uNVR for other nuisance regions (NR) suffixes WM, GM and CSF will be added accordingly. If we were simply to include the time series of all the voxels within the (NR) in our design matrix an equally large number of degrees of freedom would be lost in the analysis. Furthermore, the time series is expected to become highly colinear. One technique to deal with these problems is to include only a subspace of the extracted time series using for example PCA. Other latent variable models such as ICA or FA could be used to obtain this subspace. However, these methods often rely on an initial PCA decomposition and rotate the components within this subspace to minimise a specific cost function. In this case we are only interested in the span of the subspace and consequently PCA will suffice for this application. Before the PCA analysis it is common to normalise time series with their respective standard deviation (hence do the eigenvalue composition on the sample correlation matrix instead of the covariance matrix). This step makes small effects that are consistent (over voxels) more noticeable in the extracted components at the cost of removing information about the signal dynamics. Here we investigated both procedures, however, because we only observed marginal differences for all nuisance regions we only report results for PCA decomposition using the correlation matrix.

We demonstrated this approach in Madsen and Lund (2006) later other have utilised very similar approaches for modelling physiological artefacts see for example (Behzadi *et al.*, 2007).

Testing assumptions on residuals

Three different tests were performed on the BLUS residuals (Best Linear Unbiased with a Scalar diagonal covariance matrix) from the analyses using Statistical Parametric Mapping diagnosis (SPMd) .

- Test for arbitrary order temporal correlations based on linearity of the cumulative power spectrum (Dependence).
- Durbin-Watson test statistic the uniformly most powerful test for presence

of first order correlations (Correlation).

- Shapiro-Wilk test for normality (Normality).

For complete references see Luo and Nichols (2003). The three tests address correlation and normality on a mass univariate (voxel-wise) basis, to asses the assumptions on all voxels we calculated the number of rejections of the null hypothesis at a specific level and normalised with the nominal value.

Data

The functional datasets each consisted of 381 volumes and were acquired on a 3T (Siemens Magnetom Trio) scanner using the standard birdcage head coil. A total of 16 datasets were collected from 8 different normal subjects using a echo planar imaging gradient echo sequence with 42 slices acquired in interleaved order with the following acquisition parameters: TR=2370 ms, TE 30 ms, flip angle (FA) 90 degrees, field of view (FOV) 192x192 mm, 64x64 acquisition matrix. For delineation of anatomical regions a high resolution anatomical scan were obtained using a MPRAGE sequence with 192 sagital slices and 1 mm isotropic resolution. Additional sequence parameters were as follows: TR=15.4 ms, TE=3.93 ms, FA=9 degrees, FOV=256x256. For the duration of the scan the subjects were stimulated visually (8 Hz reversing checkerboard (expanding ring (R) and rotating wedge (W)). Each rotation/expansion lasted 30 seconds.

Prior to data analysis volume-wise 6 parameter rigid body re-alignment was performed using SPM2, and the resulting images where subsequently analysed using several different GLMs. All models were estimated using SPM2 with residuals assumed to be i.i.d. except for the 'SPM2-AR(1)' model where the standard SPM2 pre-whitening procedure using an AR(1) model was applied (Friston *et al.*, 2002). The mass univariate linear regression model applied here reads

$$\boldsymbol{y}_v = \boldsymbol{X}\boldsymbol{\beta}_v + \boldsymbol{\varepsilon}_v, \tag{5.16}$$

Where $\boldsymbol{y}_v$ is a vector containing the recorded fMRI signal at voxel v, $\boldsymbol{X}$ is the design matrix and $\boldsymbol{\varepsilon}_v$ is a vector of residuals at voxel v.

Results

For the uNVR approach we investigated the impact of the size of the NR on the extent of the region activated by the visual paradigm. Figure 5.11 shows

the mean number of activated voxels over sessions divided by the number of activated voxels identified using the standard NVR approach (see section 4.5.1) for each session. It is seen that there is a tendency towards an increase in the number of activated voxels for NR sizes up to around 1%, when the NR becomes larger than 2% the number of activated voxels decreases.

Examples of parametric maps of the $\mathcal{F}$-test statistic for physiological nuisance effects using the RETROICOR, uNVR, uNVR WM and uNVR CSF methods are shown in figures 5.12, 5.13. The maps are thresholded at $p<0.05$ GRF corrected. Dominant effects are seen near the major arteries e.g. circle of Willis and the medial cerebral artery for all 4 methods. Effects observed near the edges of the brain could be related to non-rigid movement due to either cardiac pulsation or respiration.

Results from the three tests for normality dependency, correlation and normality are summarised in figures 5.15 and 5.16. In figure 5.15 different versions of the uNVR method is compared to the NVR approach as well as other simple models. We see that the uNVR approach reduces the number of violations for all three test similar to or better than the NVR model.

In figure 5.15 the test for arbitrary order temporal correlations (top panel - dependency) showed that modelling the physiological noise by either the NVR or the uNVR approach greatly reduces the number of rejection of the null hypothesis. The AR(1) approach performs particularly bad in this test because it only models first order correlations. In the test for first order correlations (middle panel - correlation) the behaviour is more or less the same, however, the AR(1) approach performs quite good which is to be expected. In the test for normality (bottom panel - normality) the NVR and uNVR modelling schemes also perform better than the competition. Overall its can be inferred that modelling of physiological noise reduces the presence of correlation and non-normality in the residuals. The CompCorr method (Behzadi *et al.*, 2007) is basically the uNVR method without projecting out movement and high pass filter effects - this method is seen to perform similarly but consistently slightly worse than the uNVR approaches. uNVR methods based on other types of masks (CSF, WM) show performances similar to the mask based on mean residual error. The GM mask was included in order to enable comparison with the other methods, note that performances with respect to the three tests are similar, however this mask will reduce the amount of visual activity detected due to expected overlap with activated regions. It should be be noted that the simple model (no filter and no modelling of physiological effects) performs worse than all other methods for all three tests.

In figure 5.16, we see that including the regressors for explaining the change in respiration volume over time (RVT) as well as the cardiac-respiration interaction

(CXR) effect almost has no effect on the number of violations, suggesting that these effects are not dominant in this particular dataset. The test for arbitrary order temporal correlations (top panel - dependency) shows that the included NVR approaches greatly reduces the number of rejection of the null hypothesis both compared to high-pass filtering (HP) and the AR(1) approach which is included for reference. In the test for first order correlations (middle panel - correlation) we see the same tendency (but with AR(1) modelling performing much better). In the test for normality (bottom panel - normality) the HP method and the AR(1) method almost coincide whereas the NVR approaches perform much better. Overall we see that all the methods based on the NVR approach perform almost identically for all three tests. Furthermore, we see that changing the reference slice for the RETROICOR regressors does affect residual whiteness and normality substantially. The "NVR first, middle and last slice" model includes RETROICOR regressors based on 3 different reference slices (the first, middle and last based on the acquisition time) giving a total of 48 physiological regressors as compared to the 16 included in the NVR model, also it is seen that the tests are not affected substantially even though the power of each test is reduced due to the reduced degrees of freedom in the residual. In figure 5.17 the amount of variance of cardiac effects (identified by RETROICOR cardiac regressors) that are modelled by the uNVR approach as a function of the mask size (e.g. the amount of voxels from the residual error variance mask that was used PCA decomposition). From figure 5.17b we see that the regressors from uNVR and RETROICOR cardiac modelling are most similar at a mask size of around 1% also indicating that the choice of mask size is reasonable.

If false negatives are of primary concern which can be the case in for example pre-surgical planning the nuisance regressors can be orthogonalised effects of interest simply by filtering out the effect prior to PCA decomposition thereby providing better control for false negatives, however, at the cost of more false positives.

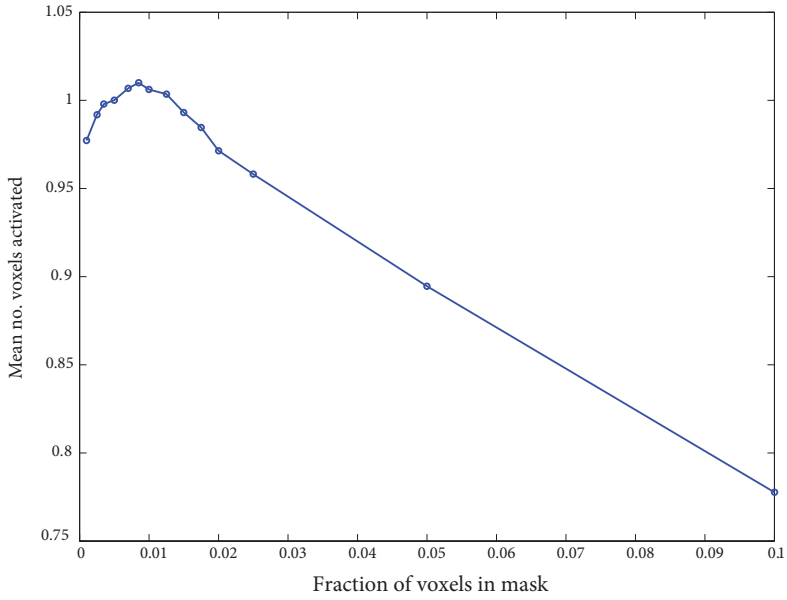


Figure 5.11: **The dependence of the mask size on the activation extent.** The figure shows the mean number of voxels activated by the visual paradigm as a function of the mask size used in the uNVR analysis. The number of voxels activated in each of the 16 sessions were determined using a threshold of $p<0.05$ (family-wise error corrected for multiple comparisons using GRF) and divided by the number of voxels detected using the NVR approach for each session. It is seen that the curve exhibits a maximum around a mask size of 1%. If the mask is smaller than this the uNVR approach may not be able to capture the dynamics of the physiological noise. However, if the mask is enlarged to above 1% the activation slowly drops presumably caused by uNVR regressors starting to explain part of the visual paradigm.

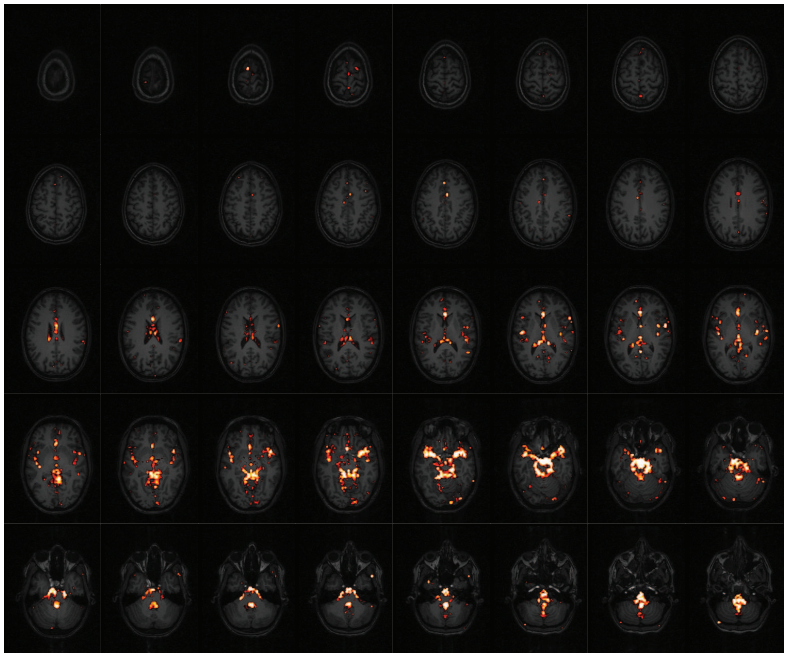


Figure 5.12: **Effects of physiological noise using supervised method.** The figure shows regions that revealed effects of the RETROICOR modelling (including both cardiac and respiratory effects). Activation was assessed by an F-test over the 16 RETROICOR regressors that entered the NVR model thresholded at $p<0.05$ FWE (GRF corrected). Effects are predominant near large vessels in the brain and in the ventricles.

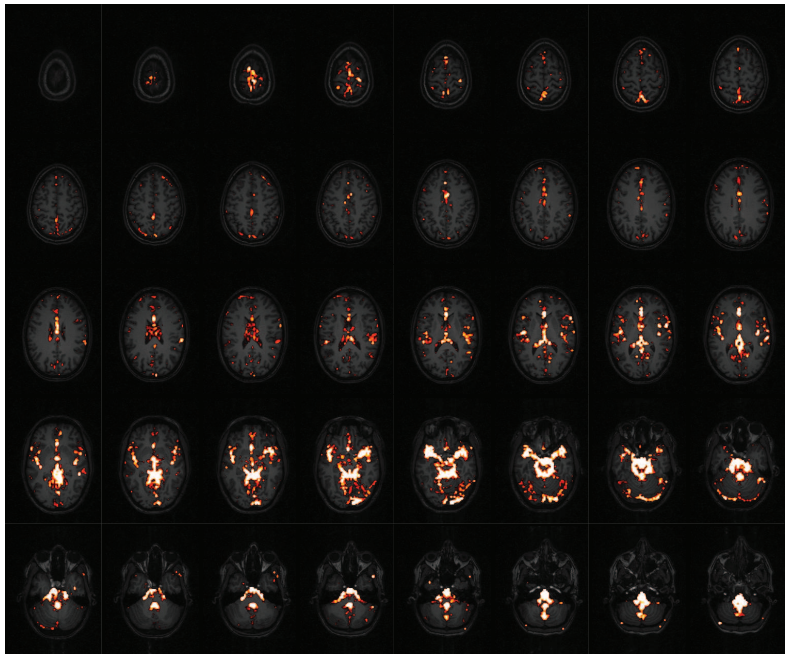


Figure 5.13: **Effects of physiological noise using unsupervised method.** The figure shows regions that revealed effects of the uNVR regressors intended to model physiological noise effects. Activation was assessed by an $\mathcal{F}$-test over the 16 physiological noise uNVR regressors that entered the uNVR model thresholded at $p<0.05$ FWE (GRF corrected). The regions show high resemblance towards regions identified by the RETROICOR method. The fact that this map shows more significant effect especially at slices near the circle of Willis is not surprising because the uNVR regressors are based on time courses from these regions. Once again effects are predominant near large vessels in the brain and in the ventricles.

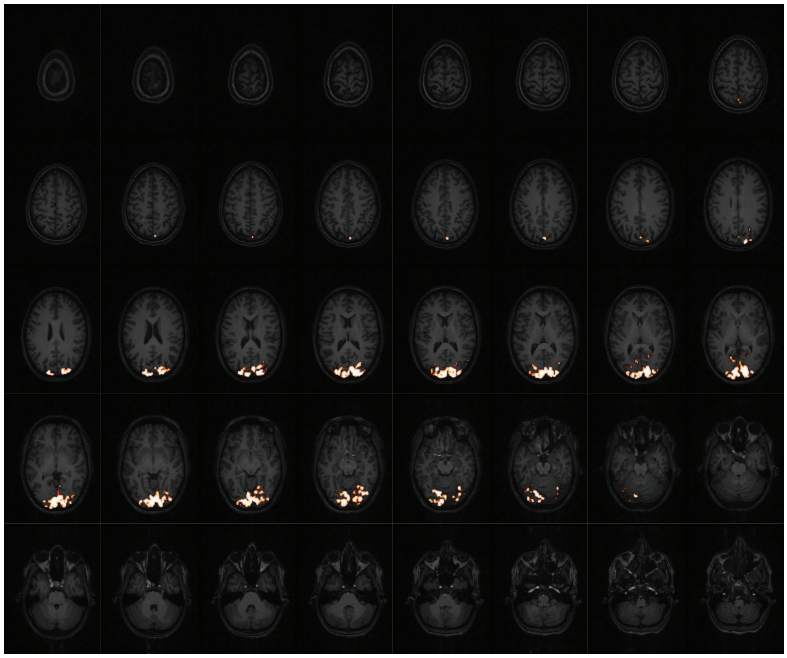


Figure 5.14: **Visual activation after unsupervised noise modelling.** The figure shows the activity of the visual paradigm assessed by an $\mathcal{F}$-test over the six regressors modelling the paradigm in the uNVR modelling approach. Activation was thesholded at a FWE of $p<0.05$ (GRF corrected). It is seen that highly significant visual activation in the occipital lobe remains after uNVR modelling. There were no obvious differences between the activation using the NVR and uNVR approach from visual inspection.

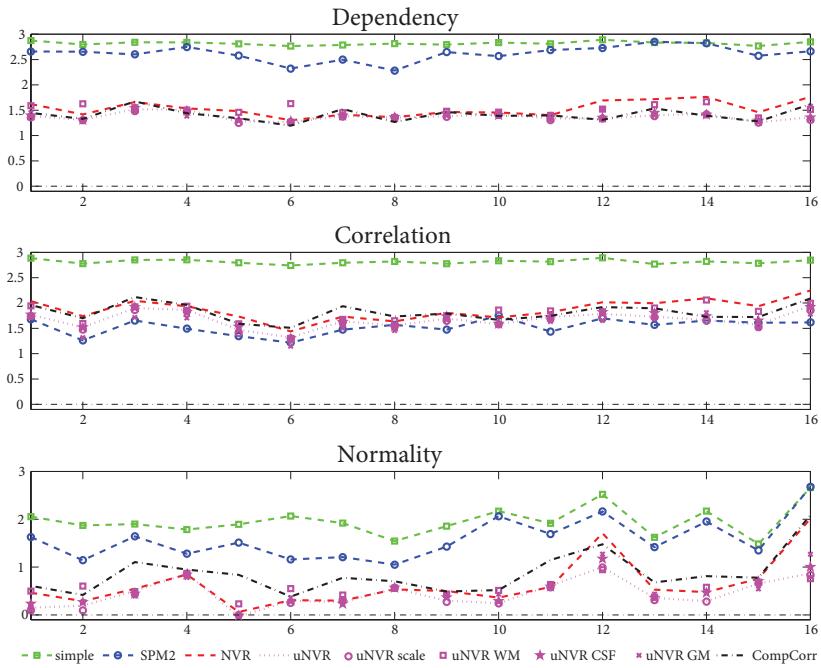


Figure 5.15: **Residual diagnostics – unsupervised vs. supervised** The three panels are related to three different tests (dependence, correlation, normality) on the assumption of uncorrelated i.i.d. residual noise. For each of the 16 sessions the ratio with which the number of rejections of the null hypothesis exceeded the expected value, a threshold of $p<0.001$ was used (e.g. it is expected that 0.1 percent of voxels would exceed this threshold by change). A value of around 1 would correspond to uncorrelated normal i.i.d. noise. The legend box indicates the symbols and lines used for each of the modelling approaches given in the bottom legend.

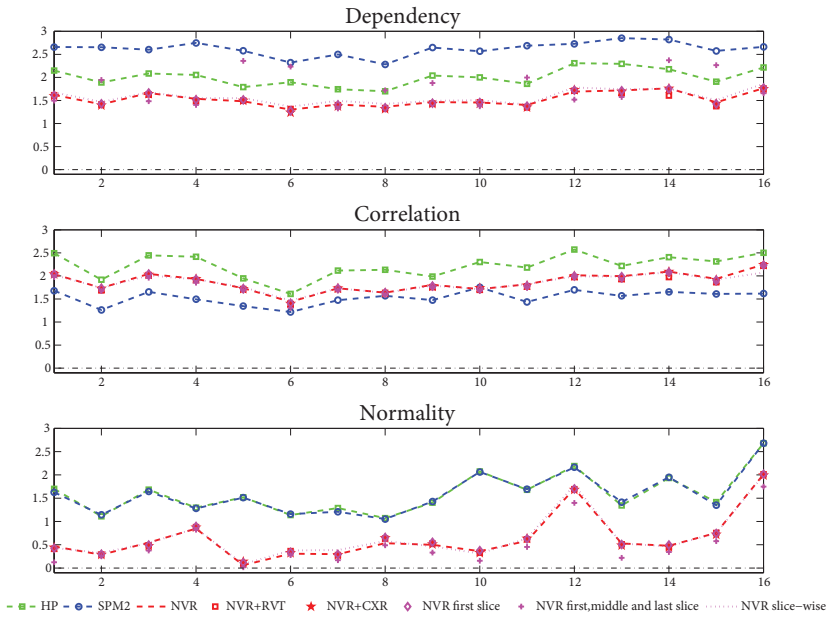
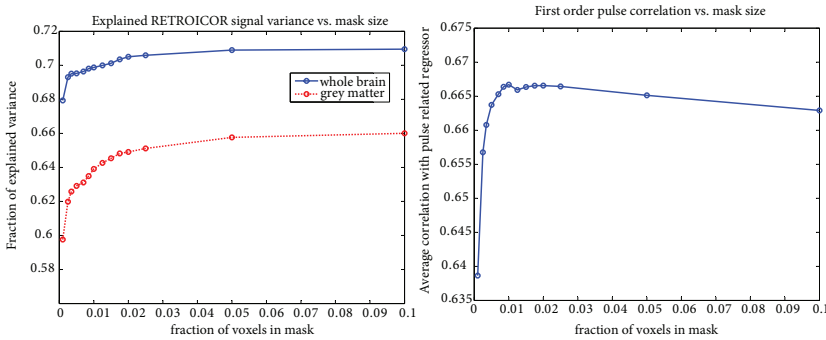


Figure 5.16: **Residual diagnostics – nuisance modelling.** The plots show how modelling of different nuisance effects influences the residual whiteness and normality. The three panels are related to three different tests (dependence, correlation, normality) on the assumption of uncorrelated i.i.d. residual noise. For each of the 16 sessions the ratio with which the number of rejections of the null hypothesis exceeded the expected value, a threshold of $p<0.001$ was used (e.g. it is expected that 0.1 percent of voxels would exceed this threshold by change). A value of around 1 would correspond to uncorrelated normal i.i.d. noise. The legend box indicates the symbols and lines used for each of the modelling approaches given in the bottom legend.



(a) Variance explained in data (b) Variance of RETROICOR explained

Figure 5.17: **Cardiac effects explained by unsupervised nuisance modelling.** The figures attempt to illustrate how the uNVR model performs in terms of modelling cardiac nuisance effects as a function of the mask size. Panel (a) shows the fraction of the variance of the physiological effects (identified by the RETROICOR regressors) that are also be explained by the uNVR regressors as a function of the uNVR mask size. A value of 1 (the maximum possible) would mean that the signal identified by the RETROICOR approach can be modelled completely by the uNVR regressors (e.g. the regressors span the same subspace). Values less than one does not indicate that the uNVR regressors explain less variance but rather that they explain different things. The blue unbroken line shows the mean fraction over sessions for the whole brain whereas the broken red line shows the same in a GM mask. The fraction of explained variance is slightly higher in the whole brain as compared to only the GM. This is to be expected because effects are dominant in areas not part of GM and because the mask is mainly constructed from these areas. For mask sizes up to around 2% the explained variance increases from 60% to around 65% and then seems to saturate at this value, a similar tendency is seen for the whole brain curve. It should however be noted that the 60% of variance explained in GM means that there is a high resemblance between the effects modelled by both schemes. Panel (b) show whether the uNVR regressors cover the subspace formed by the two first order cardiac cycle RETROICOR regressors. The explained variance reaches a maximum around a mask size of 1% and then decreases slowly.

Chapter 6

Conclusion

The availability of computational processing power continues to grow rapidly enabling more sophisticated models to be used in the processing of data. One such example is the application of Bayesian inference to neuroimaging data. However, despite these advances computationally efficient algorithms and procedures for estimation and inference remain important issues in the neuroimaging community due to the vast amounts of data collected.

The Bayesian modelling framework provides a way to incorporate prior information in the form of prior probability densities for various model parameters. Specification of prior information is particularly useful at high noise levels that often plague fMRI data. For unsupervised modelling we show how prior information in the form of non-negativity, sparseness and smoothness may enter to constrain models, improve identification and suppress noise in FA type decompositions.

Bayesian inference also facilitates comparison of different models in that the posterior probabilities are naturally penalised according to the complexity of the models (for example the number of free parameters). The main disadvantage of such practise is the computational complexity involved in the evaluation of the integrals that arise when marginalising with respect to parameters to obtain posterior probabilities of the model or certain parameters in the model. We showed that the use of conjugate priors allow these integrals to be performed

analytically, this approach may serve as a computational efficient alternative to iterative procedures such as variational Bayes and numerical integration.

Identifying signals of interest in neuroimaging experiments such as fMRI studies is a difficult task due to many sources of noise. Noise enters both in form of measurement noise and the presence of unwanted physiological contrast mechanisms such as cardiac and respiratory effects. We showed that complementary measurements of physiological signals such as the cardiac cycle enable explicit modelling of nuisance effects, and that such practise helps to bring the residual into agreement with the often simple noise model assumptions applied. In addition, we showed that unsupervised methods have great promise in identification of these nuisance effects without the need for additional measurements.

Data modelling are often based on simple linear models because parameters in these models are often easy to estimate and interpret. However, if the data does not comply with the structure enforced by these simple models the advantages are no longer present and results often become difficult to interpret. In such cases relaxing the assumption of linearity may be useful, we demonstrated this by adding modelling of delays and convolutions to factor analysis type decompositions and parallel factor analysis.

Appendix A

Equations and Computational Issues

Linear Model:

An assumption concerning the nature of reality applied unquestioningly to every relationship as though God had determined that truth must always run in straight lines.

Woodman (1979)

A.1 Probability Distributions

A.1.1 Laplace Distribution

In the Laplace distribution the probability density of the random variable, x, is based on absolute deviations (l_1-norm) from the mean/median, μ. We write this distribution $x \sim \mathcal{L}(\mu, \beta)$

$$p(x|\mu, \beta) = \frac{1}{2\beta} \exp\left[-\frac{|x - \mu|}{\beta}\right], \tag{A.1}$$

where β is a positive scale parameter. The variance is given by $2\beta^2$.

A.1.2 Multivariate Normal distribution

We write the multivariate normal (Gaussian) distribution of the $T \times T$ random vector $\boldsymbol{x}$ with mean vector $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ ($\boldsymbol{x} \sim \mathcal{N}_T(\boldsymbol{\mu}, \boldsymbol{\Sigma})$)

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{T/2} \left|\boldsymbol{\Sigma}\right|^{1/2}} \exp\left[-\tfrac{1}{2} \left(\boldsymbol{x} - \boldsymbol{\mu}\right)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right]. \tag{A.2}$$

A.1.3 Multivariate student t-distribution

We define the multivariate Student t-distribution with m degrees of freedom, covariance $\boldsymbol{\Sigma}$ and mean $\boldsymbol{\mu}$ of the random matrix $\boldsymbol{x} \sim \mathcal{T}_{T,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as

$$p(\boldsymbol{x}|\boldsymbol{\Sigma}, m) = \frac{m^{m/2} \pi^{-T/2} \left|\Sigma\right|^{-1/2} \frac{\Gamma[(m+T)/2]}{\Gamma[m/2]}}{\left(m + (\boldsymbol{x} - \boldsymbol{\mu})^\top (m - T + 1) \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)^{-\frac{m+T}{2}}},$$

$$\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}. \tag{A.3}$$

$$\mathrm{var}(\boldsymbol{x}|\boldsymbol{\Sigma}, m) = \frac{m}{m-2} \boldsymbol{\Sigma}. \tag{A.4}$$

A.1.4 Inverted Wishart

We write the inverted Wishart distribution of the random matrix Σ with scale matrix $\boldsymbol{B}$ and m degrees of freedom ($\boldsymbol{\Sigma} \sim \mathcal{W}_{T,m}^{-1}(\boldsymbol{B})$)

$$p(\boldsymbol{\Sigma}|\boldsymbol{B}, m) = \frac{\left|\boldsymbol{B}\right|^{\frac{m}{2}}}{2^{\frac{mT}{2}} \Gamma_T(\frac{m}{2}) \left|\boldsymbol{\Sigma}\right|^{\frac{m+T+1}{2}}} \exp\left[-\tfrac{1}{2} \operatorname{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right)\right], \tag{A.5}$$

where the $\Gamma_T(a)$ is the multivariate gamma function given by

$$\Gamma_T(a) = \int_{\boldsymbol{G} \in \Omega} \exp(-\operatorname{Tr} \boldsymbol{G}) \left|\boldsymbol{G}\right|^{a - \frac{(T+1)}{2}} d\boldsymbol{G} = \pi^{\frac{T(T-1)}{4}} \prod_{j=1}^{T} \Gamma(a + \tfrac{(j-1)}{2}),$$

where $\boldsymbol{\Omega}$ is the complete set of all $(T \times T)$ positive definite matrices and $\Gamma(a)$ denotes the ordinary Gamma function.

$$\mathrm{E}(\boldsymbol{\Sigma}|\boldsymbol{B}, m) = \frac{1}{m - T - 1}\boldsymbol{B}.$$

A.1.5 Normal-Inverse Wishart

We define the Normal-Inverse Wishart distribution of the random vector $\boldsymbol{x}$ of length T and $T \times T$ matrix $\boldsymbol{\Sigma}$ with the parameters $\boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\mu}$ and m as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\Sigma}|\boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\mu}, m) =& (2)^{-\frac{(m+1)T}{2}} (\pi)^{-\frac{T}{2}} |\boldsymbol{Q}|^{\frac{1}{2}} |\boldsymbol{B}|^{\frac{m}{2}} \Gamma_T(\tfrac{m}{2})^{-1} \\ & \times \exp\left[-\tfrac{m+T+2}{2} \log|\boldsymbol{\Sigma}|\right] \\ & \times \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right] \\ & \times \exp\left[-\tfrac{1}{2} \operatorname{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right)\right]. \end{aligned} \tag{A.6}$$

And introduce the notation $\mathcal{NW}^{-1}(\boldsymbol{Q}, \boldsymbol{B}, \boldsymbol{\mu}, m)$

We specifically note that the normalisation integral (the inverse of the part that does not depend on $\boldsymbol{x}$ and $\boldsymbol{\Sigma}$)

$$\begin{aligned} \int_{\boldsymbol{x} \in \mathbb{R}^T} \int_{\boldsymbol{\Sigma} \in \Omega} \exp\left[-\tfrac{m+T+2}{2} \log|\boldsymbol{\Sigma}| - \tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q} (\boldsymbol{x} - \boldsymbol{\mu}) - \tfrac{1}{2} \operatorname{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right)\right] d\boldsymbol{x} d\boldsymbol{\Sigma} \\ = (2)^{\frac{(m+1)T}{2}} (\pi)^{\frac{T}{2}} |\boldsymbol{Q}|^{-\frac{1}{2}} |\boldsymbol{B}|^{-\frac{m}{2}} \Gamma_T(\tfrac{m}{2}), \end{aligned}$$

where again $\boldsymbol{\Omega}$ denotes the complete set of all $(T \times T)$ positive definite matrices.

A.1.6 Matrix Variate Normal

We define the matrix variate normal distribution of the $T \times N$ random matrix $\boldsymbol{X} \sim \mathcal{M}_{T \times N}(\boldsymbol{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma})$

$$\begin{aligned} p(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) = & (2\pi)^{\frac{-TN}{2}} |\boldsymbol{\Omega}|^{-T/2} |\boldsymbol{\Sigma}|^{-N/2} \\ & \times \exp\left(-\frac{1}{2} \operatorname{Tr}\left[\boldsymbol{\Omega}^{-1}(\boldsymbol{X} - \mathbf{M})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{M})\right]\right), \end{aligned} \tag{A.7}$$

as a natural generalisation of the multivariate normal distribution. Here the $T \times N$ matrix $\boldsymbol{M}$ is the mean, $\boldsymbol{\Omega}$ is the $N \times N$ column covariance matrix and $\boldsymbol{\Sigma}$ is the $T \times T$ row covariance matrix. This distribution has a has a simple relation to the multivariate normal distribution $\boldsymbol{X} \sim \mathcal{M}_{T\times N}(\boldsymbol{M}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}) \Leftrightarrow \operatorname{vec} \boldsymbol{X} \sim \mathcal{N}_{TN}(\operatorname{vec} \boldsymbol{M}, \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma})$.

Simple segmentation procedure

The literature on segmentation of MRI brain volumes is extensive and remains an active research area making an detailed explanation beyond the scope of this text. However, in the following we describe a simple and crude method to identify the WM/GM border from standard anatomical MRI sequences. This border commonly serves as a basis for an unfolded representation because it is relatively well defined in T_1 weighted images.

1 Data are filtered using an edge preserving sigma filter (Kriegeskorte and Goebel, 2001; Lee, 1983) to ease the later segmentation process by reducing the number of topological defects initially present (Dale *et al.*, 1999).

2 Ventricles are located and filled; this process is done to avoid topological defects (holes) in the segmentation in these areas.

3 A brain mask is applied to remove skull and other non brain regions. This is done to avoid the region growing process to advance into these regions.

4 The white matter is found by region with a given threshold growing with a seed point known to be located inside the white matter structure.

5 Dilation and smoothing of white matter is then performed to ensure that the curvature is finite at all points.

6 Disconnection of the two hemispheres.

7 The region growing process is redone now using self-touch sensitive algorithm meaning that the outer boundaries of the region are not allowed to merge with itself and form rings. This step is performed to correct for topological errors (handles) from the segmentation in step 4. This produces a white matter segmentation free from topological defects (handles), however the result will be biased because the algorithm is only able to exclude voxels. To account for this the same process is performed for the inverse object yielding the opposite result (only inclusion of voxels). For each of the corrections the damage compared to the original segmentation

is estimated and the choice yielding the least damage is chosen (Kriegeskorte and Goebel, 2001). The resulting region is guaranteed to be free from topological errors (has Euler characteristic 2, equivalent to a sphere).

An example of the segmentation based on this procedure is displayed in figure 3.1.

A.2 Algorithms

A.2.1 Robust Gradient Method for Sparsity Regularised Problems

Using naive gradient based search used to optimise sparsity regularised problems will unfortunately often cause the solution to oscillate around zero. We will apply the following straightforward procedure of splitting the update into a reconstruction part and a sparsity part to alleviate this problem. We here aim at minimising the (LASSO) objective ($\|\boldsymbol{Y} - \boldsymbol{AS}\|_F^2 + \lambda\|\boldsymbol{S}\|_1$) with respect to $\boldsymbol{S}$), however, the procedure should generalise directly to other sparsity priors and cost functions

Algorithm A.2.1:

repeat
  Update $\boldsymbol{S}$ to minimise $\|\boldsymbol{Y} - \boldsymbol{AS}\|_F^2$
    $\boldsymbol{S}^{new} = \boldsymbol{S} - \mu(\boldsymbol{A}^\top(\boldsymbol{AS} - \boldsymbol{X})$
  Update $\boldsymbol{S}^{new}$ according to the sparsity penalty such that element crossing zero are set to zero
    $\boldsymbol{S}_{d,j}^{new} = \begin{cases} 0 & \text{if } |\boldsymbol{S}_{d,j}^{new}| < \mu\lambda \\ \boldsymbol{S}_{d,j}^{new} - \mu\lambda Sign(\boldsymbol{S}_{d,j}^{new}) & \text{otherwise} \end{cases}$
  if $\|\boldsymbol{Y} - \boldsymbol{AS}^{new}\|_F^2 + \lambda\|\boldsymbol{S}^{new}\|_1) < \|\boldsymbol{Y} - \boldsymbol{AS}\|_F^2 + \lambda\|\boldsymbol{S}\|_1)$ **then**
    $\mu = 1.2\mu$
    $\boldsymbol{S} = \boldsymbol{S}^{new}$
  else
    $\mu = \mu/2$
  end if
until convergence

A.2.2 Multiplicative Updates

Multiplicative updates is a very simple way to ensure non-negativity while minimising a cost function. We aim to minimise the cost function $C(\boldsymbol{\theta})$ of the non-negative variable $\boldsymbol{\theta}$. Further we write the gradient with respect to $\boldsymbol{\theta}$, $\nabla_{\boldsymbol{\theta}} C = [\nabla_{\boldsymbol{\theta}} C]^+ - [\nabla_{\boldsymbol{\theta}} C]^-$ where $[\nabla_{\boldsymbol{\theta}} C]^+$ and $[\nabla_{\boldsymbol{\theta}} C]^-$ denotes positive and negative parts of the gradient respectively. The multiplicative update is then given by (Lee and Seung, 1999):

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \bullet \left([\nabla_{\boldsymbol{\theta}} C]^- ./ [\nabla_{\boldsymbol{\theta}} C]^+\right) \tag{A.8}$$

Where ./ denotes element-wise division, since all quantities are non-negative, $\boldsymbol{\theta}$ will remain in the positive orthant. By adding a small constant $\varepsilon = 10^{-9}$ to both the denominator and the numerator potential division by zeros can be avoided while fixed points are left unchanged. In Lee and Seung (2000) it was proven that for the least squares the update is guaranteed to monotonically decrease the cost function.

A.2.3 Over Relaxed Bound Optimisation

The over relaxed bound optimisation as described by Salakhutdinov *et al.* (2003) is given by constructing the following simple extension of the multiplicative update

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \bullet \left([\nabla_{\boldsymbol{\theta}} C]^{-} ./ [\nabla_{\boldsymbol{\theta}} C]^{+} \right)^{\alpha}. \tag{A.9}$$

where the exponentiation is elementwise and α is a step size parameter typically tuned using a line search procedure.

A.2.4 Least Angle Regression and Selection

Consider the (LASSO) objective ($\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda \|\boldsymbol{S}\|_1$). The following algorithm commonly known as the Least Angle Regression and Selection (LARS) algorithm enables us to obtain the entire regularisation path, i.e., minimisation with respect to $\boldsymbol{\beta}$ for all values of λ at the cost of an OLS solution (Efron *et al.*, 2004).

Algorithm A.2.4 Least Angle Regression and Selection (LARS):

repeat
  $\mathbf{c} = \mathbf{X}^{\top}(\mathbf{y} - \mathbf{S}\boldsymbol{\beta})$
  $j = \arg\max(|\mathbf{c}_I|)$
  $A = [A\ j]$
  $I = I \setminus j$
  $\boldsymbol{\beta}_A = \boldsymbol{\beta}_A + \mu (\mathbf{X}^{\top}\mathbf{X})^{-1}_{A,A} \operatorname{sign}(\mathbf{c}_A)$
  $\mu = \arg\min_{\mu} \begin{cases} \exists \widetilde{\boldsymbol{\beta}}_{A_k} = 0 \text{ then } I = [I\ A_k], A = A \setminus A_k \\ \exists l \in I :\ |\widetilde{\mathbf{c}}_l| = |\widetilde{\mathbf{c}}_A| \\ \widetilde{\mathbf{c}}_A = \mathbf{0} \end{cases}$
until $\mathbf{c}_A = \mathbf{0}$

A.2.5 Non-neg. Least Angle Regression and Selection

Once again consider the (LASSO) objective ($\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\lambda\|\boldsymbol{S}\|_1$). The following algorithm allows minimisation under non-negative constrained β, i.e. $\beta \geq 0$, for all values of λ at the cost of a OLS solution.

Algorithm A.2.5 Non-negative Least Angle Regression and Selection (NLARS):

repeat
  $\mathbf{c} = \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
  $j = \arg\max(\mathbf{c}_I),\ \mathbf{c}_j > 0$
  $A = [A\ j]$
  $I = I \setminus j$
  $\boldsymbol{\beta}_A = \boldsymbol{\beta}_A + \mu(\mathbf{X}^\top\mathbf{X})^{-1}_{A,A}\mathbf{1}$
  $\mu = \arg\min_\mu \begin{cases} \exists\ \boldsymbol{\beta}_{A_k} = 0 \text{ then } I = [I\ A_k], A = A \setminus A_k \\ \exists l \in I :\ \widetilde{\mathbf{c}}_l = \widetilde{\mathbf{c}}_A \\ \widetilde{\mathbf{c}}_A = \mathbf{0} \end{cases}$
until $\mathbf{c}_A = \mathbf{0}$

Appendix B

Selected Articles

B.1 Efficient Bayesian detection of fMRI signals in temporally correlated noise

The paper concerns the use of the natural conjugate prior (Normal-Inverse Wishart) in the GLM for the analysis fMRI data. The study extends the work by Hansen *et al.* (2002) to include correlations in the prior distribution. The article was submitted to NeuroImage on the 24th of October 2007.

B.2 Shift-Invariant Multilinear Decomposition of Neuroimaging Data

The paper concerns an extension of the CP model to include components that are shifted over one mode. The model is demonstrated on simulated data as well as EEG and fMRI neuroimaging data. The article has been accepted for publication in NeuroImage.

B.3 Combining Time and Frequency Domain Optimization: Shifts, Convolution and Smoothness in Factor Analysis Type Decompositions

This paper concerns the use of mixed time and frequency domain representations in the optimisation of extended FA models. The article was submitted to Journal of Signal Processing Systems on the 15th of February 2008.

B.4 Shifted Non-Negative Matrix Factorization

The paper concern an extension of the NMF model to include factors that are delayed over one mode. The conference paper was published in the proceedings of the 2007 IEEE International Workshop on machine learning for signal processing (MLSP2007). The article for selected for oral presentation at the conference.

B.5 Shifted Independent Component Analysis

The following article concerns extension of the FA model to include shifted factors. The conference paper was published in the proceedings of the 7′th International Conference on Independent Component Analysis and Signal Separation (ICA2007).

B.6 Recovery from optic neuritis: an ROI-based analysis of LGN and visual cortical areas

The attached article was published in 'Brain', as the title suggest it concerns an ROI based visual fMRI study of patients recovering from optic neuritis. By following the examined patients over time we show how the BOLD response changes in the lateral geniculate nuclei during recovery.

B.7 Images of Illusory Motion in Primary Visual Cortex

The following article was published in 'Journal of Cognitive Neuroscience' and concerns a visual fMRI study on a perceptual filling in effect. When two separate visual stimuli are displayed sufficiently close (spatially and temporally) illusory motion between the two locations is perceived, we show that a BOLD response is observed in V1 at locations referring to the path between the two locations.

B.8 Unsupervised modelling of physiological noise artifacts in fMRI data

The following abstract concerns using a factor analysis type decomposition to eliminate effects of physiological noise in fMRI data. The abstract was selected for oral presentation at the annual meeting of the International society for magnetic resonance in medicine 2006 (ISMRM2006).

Unsupervised modelling of physiological noise artifacts in fMRI data

Kristoffer H. Madsen[1,2], Torben E. Lund[2]

[1]Informatics and Mathematical Modelling, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital, Hvidovre, Copenhagen, Denmark

Introduction:

It has been shown that including nuisance regressors to account for movement [1] and physiological artefacts [2] in Genreal Linear Models (GLM) for the analysis of fMRI data provides an efficient way to reduce correlations and non-normality in the residuals [3]. In order to model the physiological effects using the RETROICOR method [2] the respiratory and cardiac phase and frequency needs to be acquired. This however may not always be feasible. In the current study we propose a way to obtain such nuisance regressors without measures of respiratory and cardiac cycles. This is based on the observation that residual variance in fMRI data is most pronuced in regions strongly contaminated by cardiac noise [4], specifically a thresholded SPM of the residual variance in fMRI studies have remarkable resemblance towards angiograms. By capturing stucture of the time series within this region it may be possible to extact time functions that describe physiological noise in the specific data set.

Methods:

Sixteen datasets each consisting of 381 volumes of forty slices (matrix size 64x64) was acquired on a 3T scanner (Siemens Trio) using a GRE EPI sequence: voxelsize:3mm isotropic, TE=30ms, TR=2.37s. During the scanning the subject was stimulated visually (reversing checkerboard (expanding ring and rotating wedge). Each rotation/expansion lasted 30 seconds.
Following rigid body-realignment using SPM2, each dataset was subsequently analysed with six different general linear models:

"Simple" A model including baseline plus sine and cosine of the first three harmonics of the (1/30s) oscillation" "60sec-HP" a model similar to "Simple", but now including a high-pass filter modelled as a discrete cosine set with a minimum period of 60s.

"SPM2-AR(1)" A model similar to "60sec-HP" but with whitened residuals using a global AR(1) model estimated in a mask defined by the voxels where a significant effect of the paradigm was observed. (This is the recommended SPM2 procedure).

"Full" A model similar to "60sec-HP" but including several extra nuisance regressors for modelling the autocorrelation. A Volterra expansion of the movement parameters giving was used to model residual movement effects

including spin-history effects [1] (24 regressors). Respiration and cardiac noise was modelled using 16 RETROICOR [2] regressors (5 cardiac harmonics and 3 respiratory harmonics). The RETROICOR regressors is a Fourier basis spanned by the oscillations of the aliased frequencies. The cardiac frequency and phase was determined using the scanner pulsoximeter. The respiratory phase and frequency was measured using the scanner respiratory belt.

"SVD+motion" a model similar to "60sec-HP" but including the 24 regressors to model residual movement and a variable number of nuisance SVD regressors constructed using a method described later. "SVD+full" a model like "Full" but including a variable no. SVD regressors. The SVD regressors where created by fitting a model without these additional regressors and using the residual variance in this test to identify the 7.5% voxels (approx. 5000 voxels) that have the most residual variance.

Singular Value Decomposition (SVD) was then performed on a filtered version of these time series (filtered using the highpass filter and the nuisance regressors from the original model). The additional regressors is the time series corresponding to *K* largest singular values. In order to determine the number of nuisance regressors (K) we use the Laplace evidence approximation [5]. Because fMRI data is spatially correlated the number of independent observations is not equal to the number of samples (voxels) that entered the SVD therefore the number of time points was used as the number of independent observations (equivalent to pre-whitening using SVD).
After the analysis, Statistical Parametric Mapping diagnosis (SPMd) [6] was used to test the whiteness ("Dep" for arbitrary stationary dependence and "Corr" for AR(1)-type autocorrelation) and normality "Norm" of the residuals, from the different models.

Results:
The results of the SPMd of the six different analyses of the 16 different sessions and the no. component for the "SVD+motion" model are summarized in Figure 1, and for session 8 the SPMd images from the analyses are shown. When comparing the SVD regressors (from the "SVD+motion" method) it was found that on average 61% of the variance of the first harmonic cardiac RETROICOR regressors was explained by the SVD regressors. For the 2., 3., 4. and 5. harmonics the corresponding numbers were 20%, 5%, 2% and 1%. Similarly the numbers for the three respiratory harmonics were 12%, 2% and 2%. When comparing these numbers it should be noted that as the RETORICOR set is an orthogonal basis set and because the number of SVD regressors is lower than the no. of RETROICOR regressors it is impossible to explain total variance. The strong correlation between the first order harmonics should be noted as this is also where the largest effect is observed in the data.

Discussion:
In this study we show that the suggested method reduces non-normality and correlation in the residuals equivalent or better than the RETROICOR method, however without measures of the respiratory and cardiac cycles. The number of regressors used in the suggested method (3-12) is significantly lower than the no. regressors in the RETROICOR method, thus suggesting that the regressors are more efficient The obtained regressors show strong correlation towards regressors obtained using the RETROICOR method. Furthermore, the effects are in fact present in regions related to the visual paradigm even though the visual cortex is normally not regarded as a region suffering from heavy contamination by cardiac and respiratory noise.

Acknowledgements:
The Simon Spies Foundation for donating the Siemens Trio scanner.

References:
[1] Friston et al. 1996, MRM, **35**, 346-355.
[2] Glover et al. 2000, MRM, **44**, 162–7.
[3] Lund et al. 2006, NeuroImage 29, 18.
[4] Lund et al. SMRM 2001.
[5] Minka, 2000, Techinal report 514, MIT.
[6] Luo et al. 2003, NeuroImage, **19**, 1014–32.
[7] Genovese et al. 2002, NeuroImage **15**, 870-878.

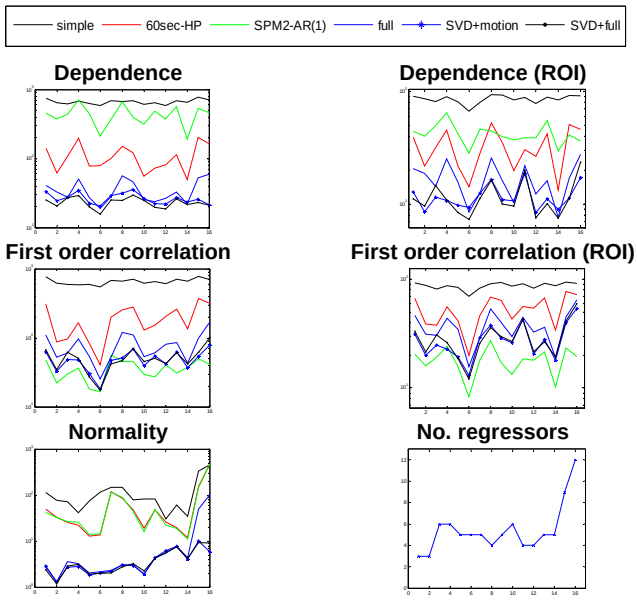


Figure 1: The three figures to the left show for the three different tests, across the 1 datasets, the factor by which rejections exceed the expected number. The different curve correspond to several different analysis of the same dataset The dataset used in Figure is from session 8. The two figures to the right show the same in a region of interest (RO defined by an F-test where the paradigm was active ($p<0.05$ FDR corrected [7]). Due t the relative few voxels in the ROI's the normality test did not have enough power in th case. The figure to the bottom right show the estimated no. of regressors for th "SVD+motion" model over the 16 datasets. The no. regressors estimated where high for sessions with more movement (session 15 and 16) thus indicating that the method able to correct more when it is needed. It is seen that the difference between th "SVD+motion" and "SVD+full" methods are only marginal.

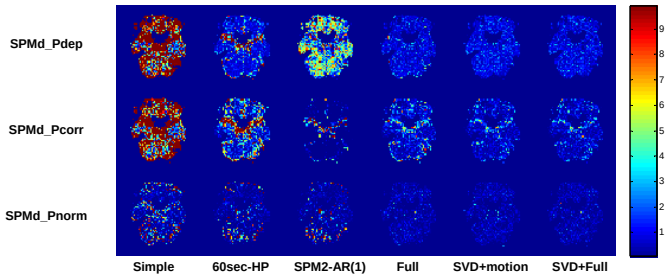


Figure 2: The figure shows the output of a SPM-diagnosis ($-\log_{10}(p)$ values) from the 6 different analyses of the same dataset (horizontal). It is seen that the three rightmost models have similar and best overall performance.

B.9 Simultaneous acquisition of polar and eccentricity mappings of the human visual cortex using fMRI

Based on a short abstract submitted to the Annual Meeting of the International society for magnetic resonance in medicine 2005 (ISMRM2005) was selected for oral presentation. In the following a poster briefly describing this research is presented.

Simultaneous acquisition of polar and eccentricity mappings of the human visual cortex using fMRI

Kristoffer H. Madsen[1,2], Torben E. Lund[1]
1 Danish Research Centre for Magnetic Resonance, Copenhagen University Hospital Hvidovre
2 Informatics and Mathematical Modelling, Technical University of Denmark
E-mail: khm@imm.dtu.dk, torbenl@drcmr.dk

Objective

In order to delimit the borders between the primary visual areas a visual field sign map is often obtained by using both a polar mapping experiment (rotating wedge) and an eccentricity mapping experiment (expanding ring) [1]. In this work we suggest showing both stimuli simultaneously to obtain both the polar and the eccentricity mapping thereby reducing the time needed to obtain a visual field sign map or alternatively improving the quality.

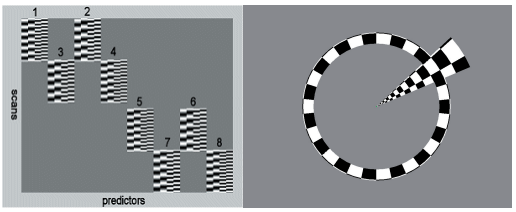


Figure 1: Experimental design
The visual stimulation consisted of a wedge rotating either clockwise (CW) or counter clockwise (CCW) at two different cycle times (time for a full rotation) on a grey background (as shown to the left). Simultaneously an expanding/contracting ring was shown also at two different cycle rates (time for one full expansion/contraction). Within both the ring and wedge a black-white checkerboard flickered at a reversal rate of 8 Hz [1], both stimuli covered a maximum of 18.4° of the subject's visual field. During the first 3 scans the subject viewed a grey screen showing a green circular fixation point covering 0.1° of the visual field; this fixation point remained on the screen for the remainder of the experiment. The stimulation then consisted of the following:
90s CCW wedge cycle rate of 25 s and expanding ring cycle rate of 30s (1 and 2)
90s CCW wedge cycle rate of 30 s and expanding ring cycle rate of 25s (3 and 4)
25s with the fixation point only
90s CW wedge cycle rate of 25 s and contracting ring cycle rate of 30s (5 and 6)
90s CW wedge cycle rate of 30 s and contracting ring cycle rate of 25s (7 and 8)
These conditions were modelled with sine and cosine predictors (1. and 2. order harmonics) with the frequency corresponding to each of the cycle rates as shown to the right.

Methods

Using a 3T scanner (Magnetom Trio, Siemens, Erlangen, Germany) and the standard birdcage head coil 525 GRE EPI volumes were acquired. The functional volumes consisted of 20 slices with 3 mm thickness oriented along the calcarine sulcus, TR=1.2 s, FOV=192 mm, 64x64 matrix, flip angle=67°. A 3D rigid body transformation was used to correct for motion. For the data analysis a general linear model (GLM) was used to model the time series of each voxel. To account for temporal auto correlation the following nuisance covariates were included in the model: High pass filter consisting of discrete cosine transform basis functions with frequencies up to 1/60 Hz, residual motion effects including spin history [3], respiratory and cardiac cycle predictors (RETROICOR) [4]. The effects of interest were modelled with harmonics of the stimulation cycle rate as described in figure 1. To determine the phase of the activation the time to peak (TTP) was found from the reconstructed signal. The hemodynamic delay was determined and accounted for by comparing the phase of stimuli running in opposite directions. In order to determine activated areas an F-test of the regressors in question was used.

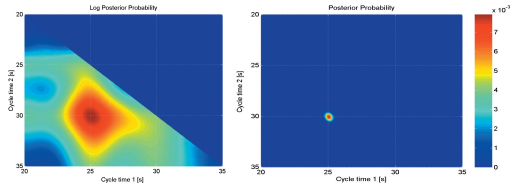


Figure 2: Detecting cycle time combinations
The figures show the mean posterior probabilities of different models in a region of interest (in the visual cortex) defined using a separate experiment. The different GLM's were constructed assuming the cycle time combinations indicated on the axes. The calculations have been performed assuming a normal inverse gamma conjugate prior. The combination 25 s and 30 s is identified as the most probable. This indicate that it is possible to resolve nearby frequencies in fMRI data. The figure to the left show the probabilities in a log colour scale whereas the more peaked figure to the right show the normalized probabilities assuming that the probability of the models tested sum to 1.

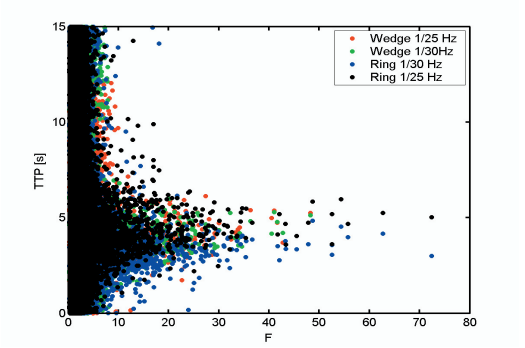


Figure 3: Hemodynamic delay
The figure show the hemodynamic delay (TTP) estimated for each voxel as a function of the F value (32,414 degrees of freedom) for each of the 4 different experiments. For higher F values the observed hemodynamic delay is both reasonable and consistent (mean 4.7 s, 0.8 s standard deviation for p<0.01 false discovery rate (FDR) [5]).

Results and Discussion

The result in figure 2 shows that discrimination between the two frequencies used is possible. Reasonable values for the hemodynamic delay were obtained as seen in figure 3. Figure 4 show phase maps. The present study used a single wedge. If independence between the right and the left hemisphere is assumed further optimization can be obtained by using a double wedge [2]. During the analysis it is assumed that the BOLD signal adds up linearly. In the present study this is assured using a relatively narrow ring and wedge (stimulation for 1.4-1.8 s). At the position of the overlap between the ring and the wedge an error is introduced as the model predicts an additive effect. However, the effect is small as this occurs very rarely for a specific position.

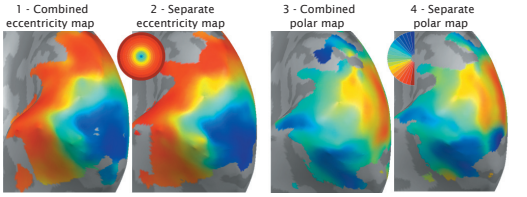


Figure 4: Phase maps
The figures show phase maps on inflated views of the right hemisphere. Figure 1 shows the eccentricity phase map from the combined experiment whereas figure 2 shows the result from a separate eccentricity mapping experiment. Likewise, figure 4 shows the result from a separate polar mapping experiment, figure 3 shows the corresponding result from the combined experiment. Note that the quality of the maps are comparable in fact the difference is close to what is normally seen as inter session variability for the separate experiments. However, the total acquisition time for the combined experiment was reduced by half as compared to the separate experiments. A FDR of 0.01 was used to threshold the maps [5].

Conclusion

The results suggest that it is possible to identify signals from a rotating wedge and an expanding/contracting ring in a single experiment thereby reducing the time needed to construct a visual field sign map. The quality of the mapping was comparable to one obtained by using two separate experiments each with the same duration.
The experimental data furthermore suggests that discrimination between two relatively closely spaced frequencies of activation is possible. This might find use in a wide range of different experiments.

References

[1] Sereno et al. 1994, Cereb Cortex, 4(6), 601-20.
[2] Slotnick et al. 2003, HBM, 18, 22-9.
[3] Friston et al. 1996, MRM, 35, 346-55.
[4] Glover et al. 2000, MRM, 44, 162-7.
[5] Genovese et al 2002, Neuroimage, 15, 870-878

Index

Bibliography

Acar, E., Aykut-Bingol, C., Bingol, H., Bro, R., and Yener, B. (2007). Multiway analysis of epilepsy tensors. *Bioinformatics*, **23**, i10–18.

Andersen, A. (1996). On the Rician distribution of noisy MRI data. *Magn Reson Med*, **36**, 331–333.

Andersen, A. H. and Rayens, W. S. (2004). Structure-seeking multilinear methods for the analysis of fmri data. *NeuroImage*, **22**, 728–739.

Andersson, C. A. and Bro, R. (1998). Improving the speed of multi-way algorithms: Part i. tucker3. *Chemometrics and Intelligent Laboratory Systems*, **42**, 93–103.

Andersson, J., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modelling geometric deformations in EPI time series. *NeuroImage*, **13**(5), 903–919.

Ashburner, J. and Friston, K. (1997). Multimodal image coregistration and partitioning - a unified framework. *NeuroImage*, **6**(3), 209–217.

Ashburner, J. and Friston, K. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, **7**(4), 254–266.

Ashburner, J. and Friston, K. (2005). Unified segmentation. *NeuroImage*, **26**, 839–851.

Ashburner, J., Andersson, J., and Friston, K. (1999). High-dimensional nonlinear image registration using symmetric priors. *NeuroImage*, **9**, 619–628.

Bandettini, P. (2007). Functional MRI today. *International Journal of Psychophysiology*, **63**, 138–145.

Bandettini, P., Wong, E., Hinks, R., Tikofsky, R., and Hyde, J. (1992). Time course epi of human brain function during task activation. *Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, **25**, 390 – 397. Cited By (since 1996): 668Export Date: 19 September 2007Source: Scopus.

Bandettini, P., Jesmanowicz, A., Wong, E., and Hyde, J. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magn Reson Med*, **30**, 161–173.

Bandettini, P., Wong, E., Jesmanowicz, A., Hinks, R., and Hyde, J. (1994). Spin-echo and gradient-echo epi of human brain activation using bold contrast: A comparative study at 1.5 t. *NMR in Biomedicine*, **7**, 12 – 20. Cited By (since 1996): 124Export Date: 19 September 2007Source: Scopus.

Beckmann, C. and Smith, S. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans Med Imaging*, **23**, 137–152.

Beckmann, C. and Smith, S. (2005). Tensorial extensions of independent component analysis for multisubject fmri analysis. *NeuroImage 25*, pages 294–311.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, **37**, 90–101.

Bell, A. J. and Sejnowski, T. J. (1995). An information maximization approach to blind source separation and blind deconvolution. *Neural Computation*, **7**, 1129–1159.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, **57**(1), 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29**(4), 1165–1188.

Birn, R. and Bandettini, P. (2005). The effect of stimulus duty cycle and "off" duration on BOLD response linearity. *Neuroimage*, **27**, 70–82.

Birn, R., Saad, Z., and Bandettini, P. (2001). Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response. *Neuroimage*, **14**, 817–826.

Blumensath, T. and Davies, M. (2004). On shift-invariant sparse coding. *International Conference on Independent Component Analysis and Blind Source Separation*, **26**, 1205–1212.

Boynton, G., Engel, S., Glover, G., and Heeger, D. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.*, **16**, 4207–4221.

Bro, R. (1998). *Multi-way Analysis in the Food Industry: Models, Algorithms & Applications*. Ph.D. thesis, Chemometrics Group, Food Technology Department of Dairy and Food Science, Royal Veterinary and Agricultural University.

Bro, R. and de Jong, S. (1997). A fast non-negativity-constrained least squares algorithm. *J. of Chemometrics*, **11**(5), 393–401.

Bro, R. and Kiers, H. A. L. (2003). A new efficient method for determining the number of components in parafac models. *Journal of Chemometrics*, **17**(5), 274–286.

Buxton, R., Wong, E., and Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magn Reson Med*, **39**, 855–864.

Buxton, R., Uludağ, K., Dubowitz, D., and Liu, T. (2004). Modeling the hemodynamic response to brain activation. *Neuroimage*, **23 Suppl 1**, S220–233.

Buxton, R. B. (2002). *Introduction to functional magnetic resonance imaging: principles & techniques*. Cambrige University Press.

Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, **18**, 23–38.

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**, 283–319.

Christensen, G. (1994). *Deformable Shape Models for Anatomy*. Ph.D. thesis, Washington University.

Christensen, R. (1996). *Plane Answers to Complex Questions*. Springer.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, **36**, 287–314.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, **9**(2), 179–194.

Darquie, A., Poline, C., Saint-Jalmes, H., and Bihan, D. L. (2001). Transient decrease in water diffusion observed in human occipital cortex during visual stimulation. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 9391–9395.

Daz-Garca, J. A. and Ramos-Quiroga, R. (2003). Generalised natural conjugate prior densities : Singular multivariate linear model. Technical Report I-03-11, Centro de Investigacon en Matemticas, Guanajuato, Gto., Mxico.

De Vos, M., Vergult, A., De Lathauwer, L., De Clercq, W., Van Huffel, S., Dupont, P., Palmini, A., and Van Paesschen, W. (2007). Canonical decomposition of ictal scalp EEG reliably detects the seizure onset zone. *Neuroimage*, **37**, 844–854.

den Dekker, A. and Sijbers, J. (2005). Implications of the Rician distribution for fMRI generalized likelihood ratio tests. *Magn Reson Imaging*, **23**, 953–959.

Deneux, T. and Faugeras, O. (2006). Using nonlinear models in fMRI data analysis: model selection and activation detection. *Neuroimage*, **32**, 1669–1689.

Donoho, D. and Stodden, V. (2004). When does nonnegative matrix factorization give a correct decomposition into parts? In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Duong, T., Yacoub, E., Adriany, G., Hu, X., Ugurbil, K., Vaughan, J., Merkle, H., and Kim, S. (2002). High-resolution, spin-echo BOLD, and CBF fMRI at 4 and 7 T. *Magn Reson Med*, **48**, 589–593.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**(2), 407–499.

Engel, S., Glover, G., and Wandell, B. (1997). Retinotopic organization in human visual cortex and the spatial precision of functional mri.

Evans, A. C., Collins, D. L., Mills, S., Brown, E. D., Kelly, R. L., and Peters, T. M. (1993). 3d statistical neuroanatomical models from 305 mri volumes. *Nuclear Science Symposium and Medical Imaging Conference*, **3**(31), 1813–1817.

Field, A. S. and Graupe, D. (1991). Topographic component (parallel factor) analysis of multichannel evoked potentials: Practical issues in trilinear spatiotemporal decomposition. *Brain Topography*, **3**(4), 407–423.

Fischl, B., Sereno, M. I., and Dale, A. M. (1999a). Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, **9**(2), 195–207.

Fischl, B., Sereno, M. I., Tootell, R. B., and Dale, A. M. (1999b). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp*, **8**(4), 272–84.

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., and Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc Natl Acad Sci U S A*, **103**(26), 10046–10051.

Frahm, J., Baudewig, J., Kallenberg, K., Kastrup, A., Merboldt, D. K., and Dechent, P. (2008). The post-stimulation undershoot in bold fmri of human brain is not caused by elevated cerebral blood volume. *NeuroImage*, **40**(2), 473–481.

Frank, O. (1930). Schatzung des schlagvolumens des menschlichen herzens auf grund der wellen und windkesseltheorie. *Zeitschrift fur Biologie (in German)*, **90**, 405–409.

Freire, L. and Mangin, J.-F. (2001). Motion correction algorithms of the brain mapping community create spurious functional activations. *Lecture Notes in Computer Science*, **2082**, 246.

Freire, L., Roche, A., and Mangin, J.-F. (2002). What is the best similarity measure for motion correction in fMRI time series? *IEEE Trans. Med. Imag.*, **21**(5), 470–484.

Friston, K., Jezzard, P., and Turner, R. (1994). Analysis of functional MRI time-series. *Human Brain Mapping*, **1**, 153–171.

Friston, K., Ashburner, J., Frith, C., Poline, J., Heather, J. D., and Frackowiak, R. (1995a). Spatial registration and normalization of images. *Human Brain Mapping*, **2**, 165–189.

Friston, K., Ashburner, J., Frith, C., Poline, J., Heather, J. D., and Frackowiak, R. (1995b). Spatial registration and normalization of images. *Human Brain Mapping*, **2**, 165–189.

Friston, K., Josephs, O., Rees, G., and Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magn Reson Med*, **39**, 41–52.

Friston, K., Mechelli, A., Turner, R., and Price, C. (2000a). Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics. *Neuroimage*, **12**, 466–477.

Friston, K. J., Holmes, A. P., Poline, J. B., Grasby, P. J., Williams, S. C., Frackowiak, R. S., and Turner, R. (1995c). Analysis of fMRI time-series revisited. *Neuroimage*, **2**(1), 45–53.

Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fmri time-series. *Magn Reson Med*, **35**(3), 346–355.

Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., and Poline, J. (2000b). To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *Neuroimage*, **12**(2), 196–208.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: theory. *Neuroimage*, **16**(2), 465–483. Comparative Study.

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, **15**, 870–878.

Glover, G. H., Li, T. Q., and Ress, D. (2000). Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. *Magn Reson Med*, **44**(1), 162–167.

Gobinet, C., Perrin, E., and Huez, R. (2004). Application of non-negative matrix factorization to fluorescence spectroscopy. *EUSIPCO*, pages 1095–1098.

Golub, G. and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, **2**(2), 205–224.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computation*. Johns Hopkins Studies in Mathematical Sciences, 3 edition.

Goutte, C., Nielsen, F., and Hansen, L. (2000). Modeling the haemodynamic response in fMRI using smooth FIR filters. *IEEE Trans Med Imaging*, **19**, 1188–1201.

Gu, H., Stein, E., and Yang, Y. (2005). Nonlinear responses of cerebral blood volume, blood flow and blood oxygenation signals during visual stimulation. *Magn Reson Imaging*, **23**, 921–928.

Gudbjartsson, H. and Patz, S. (1995). The Rician distribution of noisy MRI data. *Magn Reson Med*, **34**, 910–914.

Hansen, L. K., Nielsen, F. Å., and Larsen, J. (2002). Exploring fMRI data for periodic signal components. *Artificial Intelligence in Medicine*, **25**(1), 35–44.

Harshman, R., Hong, S., and Lundy, M. (2003a). Shifted factor analysis part i: Models and properties. *Journal of Chemometrics*, **17**, 363–378.

Harshman, R., Hong, S., and Lundy, M. (2003b). Shifted factor analysis part ii: Algorithms. *Journal of Chemometrics*, **17**, 379–388.

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**, 1–84.

Harshman, R. A. and Lundy, M. E. (1984). Data preprocessing and the extended parafac model. *In: Law, H. G., Snyder, Jr., C. W., Hattie, J. A., and McDonald, R. P. (eds.), Research Methods for Multimode Data Analysis, Praeger, New York,*, pages 216–281.

Hastie, T. and Tibshirani, R. (1990). Generalized additive models.

Henson, R., Buchel, C., Josephs, O., and Friston, K. (1999). The slice-timing problem in event-related fMRI. In *NeuroImage*, volume 9, page 125.

Hong, S. and Harshman, R. A. (2003). Shifted factor analysis part iii: N-way generalization and application. *Journal of Chemometrics*, **17**, 389–399.

Hu, X., Le, T. H., Parrish, T., and Erhard, P. (1995). Retrospective estimation and correction of physiological fluctuation in functional MRI. *Magn Reson Med*, **34**(2), 201–212.

Huettel, S. (2004). Non-linearities in the blood-oxygenation-level dependent (BOLD) response measured by functional magnetic resonance imaging (fMRI). *Conf Proc IEEE Eng Med Biol Soc*, **6**, 4413–4416.

Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128.

Jacobsen, D. J., Madsen, K. H., and Hansen, L. (2008). Bayesian model comparison in non-linear BOLD fmri hemodynamics. *Neural Computation*.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med Image Anal*, **5**, 143–156.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, **17**, 825–841.

Jin, T., Wang, P., Tasker, M., Zhao, F., and Kim, S. (2006). Source of nonlinearity in echo-time-dependent BOLD fMRI. *Magn Reson Med*, **55**, 1281–1290.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrica*, **23**, 187–200.

Kershaw, J., Ardekani, B. A., and Kanno, I. (1999). Application of Bayesian inference to fMRI data analysis. *IEEE Trans Med Imaging*, **18**(12), 1138–1153. Comparative Study.

Kershaw, J., Kashikura, K., Zhang, X., Abe, S., and Kanno, I. (2001). Bayesian technique for investigating linearity in event-related BOLD fMRI. *Magn Reson Med*, **45**, 1081–1094.

Knuth, K., Shah, A., Truccolo, W., Ding, M., Bressler, S., and Schroeder, C. (2006). Differentially variable component analysis: Identifying multiple evoked components using trial-to-trial variability. *J. Neurophysiol.*, **95**, 3257–3276.

Kriegeskorte, N. and Goebel, R. (2001). An efficient algorithm for topologically correct segmentation of the cortical sheet in anatomical mr volumes. *Neuroimage*, **14**, 329–346.

Kruskal, J. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, **18**, 95–138.

Kruskal, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, **41**(3).

Larsen, A., Madsen, K. H., Lund, T. E., and Bundesen, C. (2006). Images of illusory motion in primary visual cortex. *J. Cognitive Neuroscience*, **18**(7), 1174–1180.

Lawson, C. and Hanson, R. (1974). *Solving Least Squares Problems*, volume 15 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1995.

Le Bihan, D. (2003). Looking into the functional architecture of the brain with diffusion MRI. *Nat. Rev. Neurosci.*, **4**, 469–480.

Le Bihan, D., Urayama, S., Aso, T., Hanakawa, T., and Fukuyama, H. (2006). Direct and fast detection of neuronal activation in the human brain with diffusion MRI. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8263–8268.

Lee, D. and Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–91.

Lee, D. D. and Seung, H. S. (2000). Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562.

Lee, H., Battle, A., Raina, R., and Ng, A. Y. (2007). Efficient sparse coding algorithms. *In Proceedings of the Neural Information Processing Systems (NIPS)*, **19**.

Lee, J.-S. (1983). Digital image smoothing and the sigma filter. *Computer Vision, Graphics, and Image Processing*, **24**(2), 255–269.

Lee, J. S. L., Lee, D. D., Choi, S., Park, K. S., and Lee, D. S. (2001). Non-negative matrix factorization of dynamic images in nuclear medicine. *IEEE Nuclear Science Symposium and Medical Imaging Conference*, pages 2027–2030.

Li, T. and Song, A. (2003). Fast functional brain signal changes detected by diffusion weighted fMRI. *Magn Reson Imaging*, **21**, 829–833.

Lin, C.-J. (2007). Projected gradient methods for non-negative matrix factorization. *Neural Computation*, **19**, 2756–2779.

Liu, H. and Gao, J. (2000). An investigation of the impulse functions for the nonlinear BOLD response in functional MRI. *Magn Reson Imaging*, **18**, 931–938.

Logothetis, N. (2002). The neural basis of the blood-oxygen-level-dependent functional magnetic resonance imaging signal. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **357**, 1003–1037.

Logothetis, N., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature*, **412**, 150–157.

Lu, H., Golay, X., Pekar, J., and Van Zijl, P. (2003). Functional magnetic resonance imaging based on changes in vascular space occupancy. *Magn Reson Med*, **50**, 263–274.

Lund, T., Nørgaard, M., Rostrup, E., Rowe, J., and Paulson, O. (2005). Motion or activity: their role in intra- and inter-subject variation in fMRI. *Neuroimage*, **26**, 960–964.

Lund, T., Madsen, K., Sidaros, K., Luo, W., and Nichols, T. (2006). Non-white noise in fMRI: does modelling have an impact? *Neuroimage*, **29**, 54–66.

Lund, T. E. and Hanson, L. G. (1999). Physiological noise correction in fMRI using vessel time-series as covariates in a general linear model. In *Proceedings of the 9th Annual Meeting of ISMRM*, page 22.

Lund, T. E. and Larsson, H. B. W. (1999). Spatial distribution of low-frequency noise in fMRI. In *Proceedings of the 7th Annual Meeting of ISMRM*, page 1705.

Luo, W.-L. and Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *Neuroimage*, **19**(3), 1014–1032.

Madsen, K. H. and Hansen, L. K. (2007). Efficient bayesian detection of fmri signals in temporally correlated noise, submitted. *Neuroimage*.

Madsen, K. H. and Hansen, L. K. (2008). Combining time and frequency domain optimization: Shifts, convolution and smoothness in factor analysis type decompositions, submitted. *Journal of Signal Processing Systems*, **special issue on MLSP2007**.

Madsen, K. H. and Lund, T. E. (2006). Unsupervised modelling of physiological noise artifacts in fmri data. In *Proc. International Society of Magnetic Resonance In Medicine - ISMRM 2006, Seattle, Washington, USA*. ISMRM.

Mandeville, J., Marota, J., Ayata, C., Zaharchuk, G., Moskowitz, M., Rosen, B., and Weisskoff, R. (1999). Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J. Cereb. Blood Flow Metab.*, **19**, 679–689.

McKeown, M. and Sejnowski, T. (1998). Independent component analysis of fMRI data: examining the assumptions. *Hum Brain Mapp*, **6**, 368–372.

McKeown, M., Makeig, S., Brown, G., Jung, T., Kindermann, S., Bell, A., and Sejnowski, T. (1998a). Analysis of fMRI data by blind separation into independent spatial components. *Hum Brain Mapp*, **6**, 160–188.

McKeown, M., Jung, T., Makeig, S., Brown, G., Kindermann, S., Lee, T., and Sejnowski, T. (1998b). Spatially independent activity patterns in functional MRI data during the stroop color-naming task. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 803–810.

McKeown, M. J., Hansen, L. K., and Sejnowsk, T. J. (2003). Independent component analysis of functional mri: what is signal and what is noise? *Curr Opin Neurobiol*, **13**(5), 620–9.

Miller, K., Luh, W., Liu, T., Martinez, A., Obata, T., Wong, E., Frank, L., and Buxton, R. (2001). Nonlinear temporal dynamics of the cerebral blood flow response. *Hum Brain Mapp*, **13**, 1–12.

Miwakeichi, F., Martínez-Montes, E., Valdés-Sosa, P. A., Nishiyama, N., Mizuharam, H., and Yamaguchi, Y. (2004). Decomposing eeg data into space time frequency components using parallel factor analysis. *NeuroImage*, **22**(3), 1035–1045.

Möcks, J. (1988). Topographic components model for event-related potentials and some biophysical considerations. *IEEE Trans. Biomed. Eng.*, **35**, 482–484.

Molgedey, L. and Schuster, H. (1994). Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, **72**(34), 3634–3637.

Mørup, M., Hansen, L. K., Herrmann, C. S., Parnas, J., and Arnfred, S. M. (2006). Parallel factor analysis as an exploratory tool for wavelet transformed event-related eeg. *NeuroImage*, **29**(3), 938–947.

Mørup, M., Madsen, K. H., and Hansen, L. K. (2007a). Shifted independent component analysis. In *ICA2007*, pages 89–96.

Mørup, M., Madsen, K. H., and Hansen, L. K. (2007b). Shifted non-negative matrix factorization. In *MLSP 2007*.

Mørup, M., Madsen, K. H., and Hansen, L. K. (2008a). Approximate l0 constrained non-negative matrix and tensor factorization. In *Accepted ISCAS 2008 special session on Non-negative Matrix and Tensor Factorization and Related Problems*.

Mørup, M., Hansen, L., Arnfred, S., Lim, L.-K., and Madsen, K. (2008b). Shift invariant multilinear decomposition of neuroimaging data. *submitted NeuroImage*.

Ogawa, S., Lee, T. M., Kay, A. R., and Tank, D. W. (1990a). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A*, **87**(24), 9868–72.

Ogawa, S., Lee, T. M., Nayak, A. S., and Glynn, P. (1990b). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn Reson Med*, **14**(1), 68–78.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proc Natl Acad Sci U S A*, **89**(13), 5951–5.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609.

Osborne, M., Presnell, B., and Turlach, B. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, **20**(3), 389–403.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**(2), 111–126.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(6), 559–572.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational Bayesian inference for fMRI time series. *Neuroimage*, **19**(3), 727–741.

Petersen, N. V., Jensen, J. L., Burchardt, J., and Stødkilde-Jørgensen, H. (1998). State space models for physiological noise in fMRI time series. *NeuroImage*, **7**(4), 592.

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequetist Methods of Inference*. Robert E. Krieger Publishing Company, Malabar, Florida.

Price, R. (1763). An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, f. r. s. communicated by mr. price, in a letter to john canton, a. m. f. r. s.", philosophical transactions, giving some account of the present undertakings, studies and labours of the ingenious in many considerable parts of the world. *Philosophical Transactions of the Royal Society of London*, **53**, 370–418.

Raj, D., Anderson, A. W., and Gore, J. C. (2001). Respiratory effects in human functional magnetic resonance imaging due to bulk susceptibility changes. *Phys Med Biol*, **46**(12), 3331–3340.

Robinson, P., Drysdale, P., Van der Merwe, H., Kyriakou, E., Rigozzi, M., Germanoska, B., and Rennie, C. (2006). BOLD responses to stimuli: dependence on frequency, stimulus form, amplitude, and repetition rate. *Neuroimage*, **31**, 585–599.

Rowe, D. (2005). Modeling both the magnitude and phase of complex-valued fMRI data. *Neuroimage*, **25**, 1310–1324.

Salakhutdinov, R., Roweis, S., and Ghahramani, Z. (2003). On the convergence of bound optimization algorithms. In *Proceedings of the 19th Annual Conference on Uncertainty in Artificial Intelligence (UAI-03)*, pages 509–516, San Francisco, CA. Morgan Kaufmann Publishers.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, **2**(6), 110–114.

Schmitt, F., Stehling, M., and Turner, R. (1998). *Echo-planar imaging: theory, technique, and application*. Springer.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., and Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, **268**(5212), 889–93.

Shaobing, S. C. and Donoho, D. (1994). Basis pursuit. *28th Asilomar conf. Signals, Systems Computers*.

Siddiqui, M. (1958). On the inversion of the sample covariance matrix in a stationary autoregressive process. *The Annals of Mathematical Statistics*, **29**, 585–588.

Sidiropoulos, N. D. and Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, **14**, 229–239.

Sijbers, J., den Dekker, A., Scheunders, P., and Van Dyck, D. (1998). Maximum-likelihood estimation of Rician distribution parameters. *IEEE Trans Med Imaging*, **17**, 357–361.

Sjöstrand, K., Lund, T. E., Madsen, K. H., and Larsen, R. (2006). Sparse PCA, a new method for unsupervised analyses of fmri data. In *Proc. International Society of Magnetic Resonance In Medicine - ISMRM 2006, Seattle, Washington, USA*, Berkeley, CA, USA. ISMRM.

Smaragdis, P. and Brown, J. C. (2003). Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180.

Smith, A., Lewis, B., Ruttimann, U., Ye, F., Sinnwell, T., Yang, Y., Duyn, J., and Frank, J. (1999). Investigation of low frequency drift in fMRI signal. *Neuroimage*, **9**, 526–533.

Song, A., Wong, E., Tan, S., and Hyde, J. (1996). Diffusion weighted fMRI at 1.5 T. *Magn Reson Med*, **35**, 155–158.

Song, A., Harshbarger, T., Li, T., Kim, K., Ugurbil, K., Mori, S., and Kim, D. (2003). Functional activation using apparent diffusion coefficient-dependent contrast allows better spatial localization to the neuronal activity: evidence using diffusion tensor imaging and fiber tracking. *Neuroimage*, **20**, 955–961.

Spearman, C. (1904). "general intelligence," objectively determined and measured. *American Journal of Psychology*, **15**, 201–293.

Speck, O., Hennig, J., and Zaitsev, M. (2006). Prospective real-time slice-by-slice motion correction for fMRI in freely moving subjects. *MAGMA*, **19**, 55–61.

Stegeman, A. (2007). Degeneracy in candecomp/parafac and indscal explained for several three-sliced arrays with a two-valued typical rank. *Psychometrika*, **72**(4), 601–619.

Syskind, M. P., Larsen, J., Kjems, U., and Parra, L. C. (2007). A survey of convolutive blind source separation methods. *Springer Handbook on Speech Processing and Speech Communication*.

Talairach, J. and Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain: 3-Dimensional Proportional System - an Approach to Cerebral Imaging*. Thieme Medical Publishers, New York.

Thesen, S., Heid, O., Mueller, E., and Schad, L. (2000). Prospective acquisition correction for head motion with image-based tracking for real-time fMRI. *Magn Reson Med*, **44**, 457–465.

Thomas, C. G., Harshman, R. A., and Menon, R. S. (2002). Noise reduction in bold-based fmri using component analysis. *Neuroimage*, **17**(3), 1521–37.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Tomasi, G. (2006). *Practical and computational aspects in chemometric data analysis*. Ph.D. thesis, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark.

Truccolo, W., Knuth, K., Shah, A., Bressler, S., Schroeder, C., and Ding, M. (2003). Estimation of single-trial multicomponent ERPs: differentially variable component analysis (dVCA). *Biol Cybern*, **89**, 426–438.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, **31**, 279–311.

Van de Moortele, P., Cerf, B., Lobel, E., Paradis, A., Faurion, A., and Le Bihan, D. (1997). Latencies in fMRI time-series: effect of slice acquisition order and perception. *NMR Biomed*, **10**, 230–236.

Vaughan, J., Garwood, M., Collins, C., Liu, W., DelaBarre, L., Adriany, G., Andersen, P., Merkle, H., Goebel, R., Smith, M., and Ugurbil, K. (2001). 7T vs. 4T: RF power, homogeneity, and signal-to-noise comparison in head images. *Magn Reson Med*, **46**, 24–30.

Walker, G. (1931). On periodicity in series of related terms. *Proceedings of the Royal Society of London*, **131**, 518–532.

Wan, X., Riera, J., Iwata, K., Takahashi, M., Wakabayashi, T., and Kawashima, R. (2006). The neural basis of the hemodynamic response nonlinearity in human primary visual cortex: Implications for neurovascular coupling mechanism. *Neuroimage*, **32**, 616–625.

Wang, K., Begleiter, H., and Porjesz, B. (2000). Trilinear modeling of event-related potentials. *Brain Topography*, **12**(4), 263–271.

Warnking, J., Dojat, M., Guérin-Dugué, A., Delon-Martin, C., Olympieff, S., Richard, N., Chéhikian, A., and Segebarth, C. (2002). fMRI retinotopic mapping–step by step. *Neuroimage*, **17**, 1665–1683.

Weisskoff, R. M., Baker, J., Belliveau, J., Davis, T. L., Kwong, K. K., Cohen, M., and Roseysen, B. R. (1993). Power spectrum analysis of functionally weigthed MR data: What's in the noise. In *Proc., SMRM New York*, volume 12, page 7.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, **34**(1/2), 28–35.

Woodman, R. W. (1979). A devil's dictionary of behavioral science research terms. *The Academy of Management Review*, **4**(1), 93–94.

Woods, R., Cherry, S., and Mazziotta, J. (1992). Rapid automated algorithm for aligning and reslicing PET images. *J Comput Assist Tomogr*, **16**, 620–633.

Woods, R., Grafton, S., Holmes, C., Cherry, S., and Mazziotta, J. (1998a). Automated image registration: I. General methods and intrasubject, intramodality validation. *J Comput Assist Tomogr*, **22**, 139–152.

Woods, R., Grafton, S., Watson, J., Sicotte, N., and Mazziotta, J. (1998b). Automated image registration: II. Intersubject validation of linear and nonlinear models. *J Comput Assist Tomogr*, **22**, 153–165.

Woolrich, M., Behrens, T., and Smith, S. (2004). Constrained linear basis sets for HRF modelling using Variational Bayes. *Neuroimage*, **21**, 1748–1761.

Worsley, K. (2005). Spatial smoothing of autocorrelations to control the degrees of freedom in fMRI analysis. *Neuroimage*, **26**, 635–641.

Worsley, K., Marrett, S., Neelin, P., Vandal, A. C., Friston, K., and Evans, A. C. (1996). A unified statistical approach for determining significant voxels in images of cerebral activation. *Human Brain Mapping*, **4**, 58–73.

Worsley, K., Liao, C., Aston, J., Petre, V., Duncan, G., Morales, F., and Evans, A. (2002). A general statistical analysis for fMRI data. *Neuroimage*, **15**, 1–15.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fMRI time-series revisited–again. *Neuroimage*, **2**(3), 173–181. Comment.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London*, **226**, 267–298.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*, **20**, 45–57.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.